ROCLING 2025

The 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)

Proceedings of the Conference

November 20-22, 2025

The ROCLING organizers gratefully acknowledge the support from the following sponsors.

Sponsors

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-379-1

Message from the Conference Chairs

It is our great pleasure to welcome you to the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), held at National Taiwan University in Taipei, Taiwan, November 20 to 22, 2025.

This year's conference continues the tradition of serving as a premier forum for presenting cutting-edge research and showcasing innovative systems and techniques across the broad fields of natural language processing and speech processing.

This year marks a major milestone for ROCLING with the introduction of a dual-track submission system featuring both archival and non-archival papers. This new format provides authors with greater flexibility to present their research at different stages, combining rigorous review with opportunities for early idea exchange and community feedback.

ROCLING 2025 also launches its first Round Table Forum, a special event that brings together experienced researchers, industry experts, early-career scholars, and students for in-depth small-group discussions. The forum offers a valuable opportunity to exchange ideas across disciplines, receive constructive feedback, and build lasting professional connections.

We are also proud to note that this year's Program Committee is chaired by a team of outstanding young scholars from Taiwan. Their vision and dedication have shaped a high-quality, forward-looking program that reflects the energy, creativity, and diversity of our growing research community.

The conference program features two keynote speeches delivered by world-renowned scholars, who will share their insights into the future of language understanding and speech generation, two tutorials addressing AI-driven hearing assistive technologies and audio intelligence, and three special sessions exploring key directions in persuasive language in the age of AI, sentiment and medical text analysis, and speech recognition for Taiwanese Hakka languages.

We thank all authors, reviewers, organizers, and volunteers, as well as our sponsors, for their dedicated contributions. We hope that ROCLING 2025 will inspire new ideas, foster lasting collaborations, and strengthen our shared mission to advance human language and speech technologies.

Warm regards,

Prof. Yun-Nung Chen, National Taiwan University
Prof. Hung-Yi Lee, National Taiwan University
Prof. Pu-Jen Cheng, National Taiwan University

Conference Chairs, ROCLING 2025

Organizing Committee

Conference Chairs

Yun-Nung Chen, National Taiwan University
Hung-Yi Lee, National Taiwan University
Pu-Jen Cheng, National Taiwan University

Program Chairs

Kai-Wei Chang, Massachusetts Institute of Technology
Ke-Han Lu, National Taiwan University
Chih-Kai Yang, National Taiwan University
Zhi-Rui Tam, National Taiwan University
Wen-Yu Chang, National Taiwan University
Chung-Che Wang, National Taiwan University

Publication Chair

Kai-Wei Chang, Massachusetts Institute of Technology

Tutorial Chair

Lung-Hao Lee, National Yang Ming Chiao Tung University

Special Session Chairs

Tzu-Mi Lin, National Yang Ming Chiao Tung University
Siaw-Fong Chung, National Chengchi University
Yuan-Fu Liao, National Yang Ming Chiao Tung University

Registration Chair

Yu Tsao, Academia Sinica

Web Chairs

Yik Ming Chin, National Taiwan University
Jui-Wei Fu, National Taiwan University
Wei-Tang Hsu, National Taiwan University

Program Committee

Program Committee

Jason S. Chang, National Tsing Hua University
Kai-Wei Chang, Massachusetts Institute of Technology
Wen Yu Chang, National Taiwan University
Yung-Chun Chang, Taipei Medical University
Chien-Cheng Chen, National Taiwan University
Chien-Hung Chen, National Taiwan University
Chung-Chi Chen, National Institute of Advanced Industrial Science and Technology
Kuan-Yu Chen, National Taiwan University of Science and Technology
Pin-Er Chen, National Taiwan University
Po-Chun Chen, National Taiwan University
Wei-Chih Chen, National Taiwan University
Jou-An Chi, National Taiwan University
Huang-Cheng Chou, National Tsing Hua University
Siaw-Fong Chung, National Chengchi University
Min-Yuh Day, National Taipei University
Yao-Chung Fan, National Chung Hsing University
Bo-Han Feng, National Taiwan University
Chi-Yuan Hsiao, National Taiwan University
Chan-Jan Hsu, MediaTek Research
En-Pei Hu, National Taiwan University
Chao-Shih Huang, National Yang Ming Chiao Tung University
Chien-Yu Huang, Carnegie Mellon University
Hen-Hsen Huang, Academia Sinica
Hsiao-Ying Huang, National Taiwan University
Sung-Feng Huang, NVIDIA
Wei Ping Huang, National Taiwan University
Chih-Chung Kuo, National Yang Ming Chiao Tung University
Lung-Hao Lee, National Yang Ming Chiao Tung University
Yu-Ting Lee, National Chengchi University
Yuan-Fu Liao, National Yang Ming Chiao Tung University
Claire Lin, National Taiwan University
Tzu-Quan Lin, National Taiwan University
Yi-Cheng Lin, National Taiwan University
Yu-Xiang Lin, National Taiwan University
Chi-Liang Liu, Kotoba Technologies
Ke-Han Lu, National Taiwan University
Wenze Ren, National Taiwan University
Ming-Hsiang Su, Soochow University
Zhi Rui Tam, National Taiwan University
Yu-Chien Tang, National Yang Ming Chiao Tung University
Ming-Feng Tsai, National Chengchi University
Hou-Chiang Tseng, National Taiwan University of Science and Technology
Wei-Cheng Tseng, University of Texas at Austin
Yuan Tseng, The University of Texas at Austin
Chung-Che Wang, National Tsing Hua University
Hsin-Min Wang, Academia Sinica

Jenq-Haur Wang, National Taipei University of Technology
Jun-You Wang, National Taiwan Normal University
Jheng-Long Wu, Soochow University
Chih-Kai Yang, National Taiwan University
Pei-Ying Yang, National Taiwan University
Wen-Chao Yeh, National Tsing Hua University
An-Zi Yen, National Yang Ming Chiao Tung University
Liang-Chih Yu, Yuan Ze University

Keynote Talk
Towards Social Agents

Asli Celikyilmaz
Meta FAIR

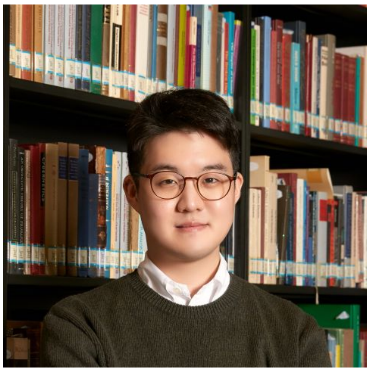
November 21st, 2025 – Time: **09:10 - 10:10** – Room: **2F, Space M Session Room / 1F, R117**

Abstract: As language models evolve into social agents, new challenges arise around reasoning, collaboration, and understanding others' minds. I'll share three directions that move us beyond next-word prediction and towards social agents: collaborative reasoning, where agents learn to communicate, coordinate, and build on each other's ideas; mental modeling (theory of mind), the ability to represent what others know, believe, or intend; and social preference alignment, where models learn human values not just from isolated feedback but through extended, context-rich interaction. Together, these efforts aim to build agents that not only converse fluently but also reason jointly, interpret intentions, and evolve toward more adaptive, long-horizon social intelligence.

Bio: Asli Celikyilmaz is a Senior Research Manager at Fundmendals AI Research (FAIR). Formerly, she was Senior Principal Researcher at Microsoft Research (MSR) in Redmond, Washington. She is also an Affiliate Associate Member at the University of Washington. She has received Ph.D. Degree in Information Science from University of Toronto, Canada, and later continued her Postdoc study at Computer Science Department of the University of California, Berkeley. Her research interests are mainly in deep learning and natural language, specifically on language generation with long-term coherence, language understanding, language grounding with vision, and building intelligent agents for human-computer interaction She is serving on the editorial boards of Transactions of the ACL (TACL) as area editor and Open Journal of Signal Processing (OJSP) as Associate Editor. She has received several "best of" awards including NAFIPS 2007, Semantic Computing 2009, CVPR 2019, EMNLP 2023.

Keynote Talk
Giving Voice and Face to AI

Joon Son Chung
KAIST



November 22nd, 2025 – Time: **09:10 - 10:10** – Room: **2F, Space M Session Room / 1F, R117**

Abstract: As AI systems advance, building natural and intuitive multimodal interfaces is becoming increasingly critical. This talk examines technologies that equip AI with both a voice and a face, improving their capacity for seamless, expressive communication with humans. We will discuss how incorporating visual and linguistic signals into speech synthesis enables alignment between acoustic output, facial and textual attributes, yielding more natural and expressive speech generation. Our recent work synthesises speech directly from visual inputs, enabling communication where audio signals are limited or absent. In parallel, we present our talking head synthesis system, where audio inputs generate lifelike facial animations, effectively giving a face to the AI's voice and enriching the multimodal interaction.

Bio: Joon Son Chung is an associate professor at the School of Electrical Engineering, KAIST, where he is directing the Multimodal AI Lab. Previously, he was a research team lead at Naver Corporation, where he managed the development of speech recognition models for various applications including Clova Note and LINE CLOVA AI Speaker. He received his BA and PhD from the University of Oxford, working with Prof. Andrew Zisserman. His work has been published in top-tier venues such as TPAMI and IJCV, and he has received several paper awards, including at Interspeech and ACCV. His research interests include speaker recognition, multimodal learning, visual speech synthesis and audio-visual speech recognition. He is a co-author of the well-known audio-visual dataset for human speech, VoxCeleb. According to Google Scholar, his work has accumulated over 17,000 citations.

Table of Contents

應用詞嵌入技術訓練中文可聽性模型以預測口語文本難度
(Training a Chinese Listenability Model Using Word2Vec to Predict the Difficulty of Spoken Texts)

錢彥翔 Yen-Hsiang Chien[1], 曾厚強 Hou-Chiang Tseng[1], 陳冠宇 Kuan-Yu Chen[2], 宋曜廷 Yao-Ting Sung[3]

[1]Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology
[2]Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology
[3]Department of Educational Psychology and Counseling, National Taiwan Normal University
wwwww43992001@gmail.com, tsenghc@mail.ntust.edu.tw, kychen@mail.ntust.edu.tw, sungtc@ntnu.edu.tw

摘要

隨著數位學習的普及，越來越多學習者會接觸到以影音為主的教材。對於學齡前至小學低年級的學生來說，受限於有限的識字能力，更仰賴以聲音與影像為主的教材來獲取知識。早期的可讀性模型主要針對書面語設計，其在口語材料上的適用性尚待驗證。為解決此問題，本研究針對中文不同年級的口語材料，探討不同斷詞工具、語言模型，對自動分級模型效能的影響。本研究以支持向量機進行年級分級，旨在自動預測教材對應年級，協助學習者選擇適性化教材。結果顯示，詞嵌入維度較高的語言模型在分級效能上有較佳表現，準確率最高可達61%，鄰近準確率則為76%。此結果有助於未來在數位學習平台或教學資源推薦系統中，自動化地為學生選擇合適的聽力教材，提升學習的成效。

Abstract

With the proliferation of digital learning, an increasing number of learners are engaging with audio-visual materials. For preschool and lower elementary students, whose literacy skills are still limited, knowledge acquisition relies more heavily on spoken and visual content. Traditional readability models were primarily developed for written texts, and their applicability to spoken materials remains uncertain. To address this issue, this study investigates the impact of different word segmentation tools and language models on the performance of automatic grade classification models for Chinese spoken materials. Support Vector Machines were employed for grade prediction, aiming to automatically determine the appropriate grade level of learning resources and assist learners in selecting suitable materials. The results show that language models with higher-dimensional word embeddings achieved better classification performance, with an accuracy of up to 61% and an adjacent accuracy of 76%. These findings may contribute to future digital learning platforms or educational resource recommendation systems by automatically providing students with appropriate listening materials to enhance learning outcomes.

關鍵字：可聽性、預訓練語言模型、自然語言處理

Keywords:Listenability, Pre-Trained Language Models, Natural Language Processing

1 緒論

「適性學習」核心宗旨在於針對每位學習者的程度與需求，對其教學方式提供最適合建議以優化其學習效率(Fariani et al., 2023)，根據Chall (1983) 著作《閱讀發展階段》中提出，閱讀能力的成長可以分為六個階段——「閱讀準備期」(Pre-Reading)、「基礎識字與破譯期」(Initial Reading and Decoding)、「鞏固與流暢期」(Confirmation & Fluency)、「閱讀學習新知期」(Reading for Learning the New)、「多角度閱讀期」(Reading at Multiple Viewpoints)和「建構與重組期」(Construction-Reconstruction)。每個階段對應不同的年齡與能力。基於此理論，學習材料的難度選擇便格外關鍵，若材料過於簡單，學習者難以獲得新知識；若過於困難，

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 1–10
November 20-22, 2025 ©2025 Association for Computational Linguistics

則可能造成過重的認知負荷（Cognitive Load）(Cambria & Guthrie, 2010)。因此，事先評估閱讀材料的難易度並針對學習者能力來提供適切內容，將有助於提升整體學習成效(Bahmani & Farvardin, 2017)。文本可讀性（Text Readability），指文本可以被理解的程度，可讀性越高，表示文本越能夠被理解(Dale & Chall, 1949)。為了評估閱讀材料的難易度，許多學者自 20 世紀中葉起便開始研究文本可讀性，以評估書面文本難易度 (Dale & Chall, 1948; Flesch, 1948; Gunning, 1952; Kincaid et al., 1975; Mc Laughlin, 1969)。隨著時代進入 21 世紀，機器學習技術的快速發展，評估文本可讀性的方式也日益自動化與精確化，越來越多研究將文本可讀性公式計算得到的分數作為特徵，再結合其他語言特徵後來訓練分類模型，以自動評估文本可讀性。例如，Petersen and Ostendorf (2009) ，採用支持向量機（Support Vector Machine，SVM）(Cortes & Vapnik, 1995)，結合平均句法樹高度（Parse Tree）、平均名詞數、Flesch–Kincaid 分數…等等進行英語文本的閱讀等級判別。Liu et al. (2021) 則比較多種機器學習模型在線上健康資訊可讀性分級的表現，模型包括 XGBoost (Chen & Guestrin, 2016)、Random Forest (Breiman, 2001)、Bayes Net (Pearl, 2014)…等等，結果顯示，整合多個模型方法表現最佳。這些研究驗證了結合語言特徵的機器學習模型，能顯著提升文本分級的準確性。

由上述可知，許多研究在探討文本可讀性，然而，隨著時代演變，在網際網路與寬頻的快速進步下，學習型態也產生強烈改變，學習型態從以往的書面「文字為主」逐漸轉向影音拓展，如：YouTube、Podcast、有聲書，再到線上教學平台，如因材網 (教育部因材網, 2020)等，以「聲音與影像」為主的數位內容，使知識獲取的方式更加多元化。學習者不再僅透過閱讀文字，還可以透過「聆聽講述」與「觀看影像」來獲取知識。而對於年齡較小的兒童而言，他們往往尚未具備理解複雜書面文本的能力 (Hogan et al., 2014)。根據 Chall (1983) 的閱讀發展階段理論，從 6 個月大到 8 歲這段期間，處於「閱讀準備期」到「鞏固與流暢期」，也就是學齡前至小學 3 年級這一段期間，學童的閱讀能力尚在發展，識字能力有限。對於這樣的學習者，他們可能更仰賴聽覺來理解世界，甚至透過語音與歌謠式的學習，來提高學習動機與理解程度。換言之，對於學齡前至國小低年級的學生而言，在「聽」和「讀」的理解能力上，可能會出現差異 (Catts et al., 2006)；根據 Ehri and Wilce (1985) 研究發現，學齡前及初學閱讀的兒童常常能夠正確理解與說出某些詞彙，但未必能夠準確辨認這些詞彙，顯示口語理解與閱讀辨識之間存在明顯落差。因此，在學習型態改變之下，傳統的可讀性公式雖然能夠有效評估書面文本難易度，但是否合適評估學齡前及初學閱讀兒童的口語材料的難度，便成為值得關注的議題。

事實上，針對口語材料難度的分析，稱之為可聽性（Listenability）模型，並已有多項相關研究 (Alghamdi et al., 2022; Fang, 1966; Yoon et al., 2016)。可聽性指語音材料被聽者理解的程度 (Harwood & Cartier, 1952)。其評估有助於教師與學習者快速篩選適合其聽力程度的語料，進而提升學習動機與聽力成效 (Alghamdi et al., 2023)。在教學實務與相關研究中，常見的語音材料大多為教學錄音、新聞稿與線上課程，這些內容多會轉換為口語文本，作為可聽性分析與模型建構的基礎。因此，口語文本的難度評估也逐漸成為可聽性研究的重要方向。例如：在美國 Kayam (2018) 以三種可讀性公式分析 2016 年美國總統候選人的演講與訪談文本，發現語言結構較為簡單、難度較低的文本，更容易觸及廣泛的受眾。Bayona et al. (2023) 等人採用四種可讀性公式，評估線上口語教學素材的難度，結果顯示這些公式雖能在一定程度上區分語料難度，但當教材難度提高時，僅依賴可讀性公式，可能難以充分反映口語材料的真實難度。該研究也進一步指出，對於高階或不同類別的口語材料，其可聽性差異往往無法僅以文字特徵來評估，顯示出未納入語速、斷句等語音特徵，可能導致難度分級的錯誤。值得一提的是 Leal et al. (2024) 等人針對同為兒童設計的語料，兒童電影字幕與兒童非小說書面文本，使用包含 200 個語言特徵的 NILC-Metrix 工具進行比較分析，發現兒童電影字幕與兒童非小說書面文本在語言複雜度、詞彙豐富度、句法結構、語篇銜接等面向均發現顯著差異。其

中，兒童非小說書面文本具有較高的語法複雜度與詞彙多樣性，語篇結構更具連貫性，而兒童電影字幕則更趨向簡化、重複。這說明即使語料來自同一個年齡層，書面語料與口語材料在語言特徵上仍存在顯著差異。這與Louwerse et al. (2004) 等人的研究結果相似，該研究從詞彙、句法、語篇結構和凝聚性等多個語言特徵進行分析，證實當分析層次提升到語篇與凝聚性特徵時，口語和書面語之間的差異明顯。這不僅體現在語音中的語速、停頓等語音現象，更在詞彙、句法、語篇結構等語言特徵上有本質差異。由此可見，單純仰賴書面文本所開發的可讀性公式，難以直接應用於口語材料。因此，發展可聽性模型時，除了可考慮納入語音特徵外，更應重視口語材料獨有的語言特徵，以提升分級的準確性和實用性。整體而言，雖有部分研究者嘗試以可讀性公式與部分語言特徵應用於評估口語材料的難度，但大多數研究忽略了口語材料中獨有語言特徵與語音特徵，因此可能無法反映其真實難度。

隨著機器學習技術的成熟，語言模型在自然語言處理（Natural Language Processing，NLP）的任務中應用廣泛，有研究者試圖利用語言模型，如 Word2Vec (Mikolov et al., 2013)、BERT (Devlin et al., 2019)應用於文本分級任務。應用語言模型進行文本分級的可行性已被多項研究證實 (Uçar et al., 2024; Zhang, 2024)。然而，目前利用語言模型來作為特徵，以訓練出可聽性模型來評估口語文本難度的相關研究仍相對有限。因此，鑑於網路上現有的大量預訓練語言模型與豐富語料資源，本研究將採用二種不同語言模型：中央研究院 Word2Vec (Chen & Ma, 2018)、奧斯陸大學 Word2Vec (Fares et al., 2017)，訓練學齡前至高中三年級（K-12）中文口語文本可聽性模型，並比較不同模型準確率。本研究的內容如下：第二節將闡述可聽性模型相關研究，第三節講述研究設計與資料，第四節分析研究結果，第五節將總結以及未來展望。

2 相關研究

早期可聽性模型研究主要將各類語言特徵應用於口語文本來驗證模型效能，如：Rogers (1962) 收集美國 12 個年級共 480 段口語錄音為依變項，並以多元迴歸分析（Multiple Regression Analysis）對多個語言特徵作為自變項，用以預測聽眾所需的年級水平。結果顯示，句法複雜度和詞彙難度能夠有效預測年級水平，其預測年級與實際年級相差不超過兩個年級。Fang (1966) 則使用相關性分析（Correlate Analysis），探討多音節詞比例與平均句長等語言特徵，對書面新聞稿與口語新聞稿之間差異的影響。研究結果顯示，在口語新聞稿中，每句中包含越多多音節詞，會顯著增加聽眾的理解負擔。換句話說，句子中若包含越多發音較長、超過一個音節的詞，會讓內容聽起來更難懂。由上述可知，利用語言特徵來訓練可聽性模型以評估口語文本的難度具有可行性。隨著科技的發展，開始有研究者考量到口語中的語音特徵，例如 Yoon et al. (2016) 等人收集來自聽力測驗、新聞、訪談等口語樣本；先由不同程度英語學習者評定口語樣本的難度，並分為初級、中級與高級做為口語樣本難度標籤，再以語音特徵，包括語速、停頓頻率⋯等等，以及語言特徵來進行相關性分析，並選出了 12 項重要特徵建立多元迴歸模型。結果顯示，模型若僅納入語音特徵，預測準確度略高於僅使用語言特徵，而語音特徵與語言特徵同時納入時，準確度最高。Alghamdi et al. (2022) 則收集了涵蓋人文、商管、理工等領域的大學講座影片，以 EFL（English as a Foreign Language）B1 水準的學生針對每部影片的理解難度進行評分，最後取所有學生的平均分數作為該影片的難度標籤。研究在特徵方面考慮了共計 168 項語音特徵與語言特徵，包括發音比例、音節平均時長、平均停頓時長⋯等等，透過相關性與多元共線性（Multicollinearity）檢驗從 168 個特徵中選出 130 項特徵來訓練出可預測影片難度的偏最小二乘迴歸（Partial Least Squares Regression，PLS）模型。研究結果顯示，經過篩選後語音特徵與語言特徵，用於PLS模型時，在測試集上可解釋52%的總變異，表現優於只納入句長與音節長度特徵的模型。綜上所述，從早期的可聽性模型僅考慮語言特徵，到納入語音特徵，讓模型預測能力有所提升。然而，這些研究多聚焦於句長、音節數、語速、停頓等語言特徵和語音特徵。

隨著機器學習與自然語言處理技術的進步，文本可讀性研究已逐漸從分析語言特徵的方法，開始結合詞嵌入（Word Embedding）技術，例如 Word2Vec、BERT 等語言模型，以自動捕捉詞彙間的語義關聯。例如，Uçar et al. (2024) 等人的研究聚焦於多語言的科學教育文本，利用基於 BERT 與 Longformer (Beltagy et al., 2020)的深度學習模型，對英語、西班牙語及巴斯克語的科學教材進行文本可讀性評估。他們建立專門的教育語料庫，通過比較傳統機器學習方法與深度學習模型的效能，發現深度學習模型在準確率與 F1 分數上皆顯著優於傳統方法，尤其是 BERT 模型表現突出。而 Zhang (2024) 在國際漢語教學領域中開發了基於 BERT 的 RCS-CSLT 系統，並結合中國教育部頒布的「國際漢語教育中文能力分級標準」（Chinese Proficiency Grading Standards，CPGS）以及特定的中文語言特徵，包括詞彙豐富度、句法複雜度…等等。進行以漢語作為第二語言文本的可讀性分級。他們收集並標記大量國際漢語教學文本，訓練並微調 BERT 模型，結果顯示使用基礎 BERT 模型的平均準確率 84.1%，而使用結合語言特徵的 RCS-CSLT 模型後，平均準確率可提升至 89.8%。

由上述研究可知，近年來，可讀性模型已經普遍利用語言模型來進行訓練，以分析書面文本的難度。然而，相較於書面可讀性模型，現有的可聽性模型研究較少基於語言模型進行訓練。此外，對於語言模型分級任務而言，詞嵌入維度的選擇常被認為是影響分級效能的重要因素 (Yin & Shen, 2018)。一般來說，詞嵌入的維度越大，模型能夠捕捉到的語義資訊就越豐富，但同時維度過高也會導致泛化能力下降，必須在表徵能力與效能之間取得平衡 (Melamud et al., 2016)。因此，若能進一步比較不同詞嵌入維度的語言模型，對於 K-12 年級口語文本的分級任務中的效能，也將是一個值得研究的議題。綜上所述，本研究將比較不同維度語言模型對 K-12 年級口語文本分級效能的影響，期望為中文口語文本可聽性模型的建立提供實證依據與新思路。

3 研究設計

本研究的整體流程如圖 1 所示，依序為資料收集、資料前處理、特徵提取（Feature Extraction），以及分級器訓練與驗證等階段。在資料收集階段，本研究收集 2865 篇中文口語材料，涵蓋學齡前至高中三年級。其中，學齡前階段語料來自多個為 3 至 8 歲幼兒教育的 YouTub 頻道 (Peppa-Pig-Chinese-Official, n.d.; PTSKIDS, n.d.; XIAOXINGXING-樂樂 TV, n.d.; 北鼻故事屋 YouTube 頻道, n.d.)，1 至 12 年級則取自臺灣臺北酷課雲 (臺北酷課雲 Taipei CooC-Cloud, n.d.)的官方教學影片。所有影片皆由具教師資格之教師錄製，並依據課程難度編排，內容具有公信力。各年級影片數量詳見表 1。

在語音資料取得後，本研究統一採用微軟 Azure 語音辨識 (Microsoft, 2025)進行語音轉文字（Speech to Text，STT）處理，並進行人工校對，完整保留口語中的冗詞，以維持資料的原始性與真實口語特徵。在口語文本取得後，因有別於英文在語言結構上詞彙之間留有空格，中文在語言結構上詞彙之間缺乏空格作為分隔，因此需要斷詞（Word Segmentation）來劃分詞彙，做為後續輸入語言模型的前置準備。現行的斷詞工具在斷詞策略上亦有優劣，斷詞的精確度不僅影響詞彙的劃分，也會進一步影響語言模型在詞嵌入建構上的表現。因此，本研究分別採用 Jieba (fxsjy, 2012)與 CkipTagger (Li P-H, 2019)兩套中文斷詞工具進行斷詞，以評估斷詞對於詞嵌入建構與模型效能的影響。

斷詞完成後，斷詞結果將作為語言模型的輸入。本研究在語言模型部分採用 Word2Vec，而 Word2Vec 常見的兩種訓練策略為 Skip-gram 與 Continuous Bag-of-Words（CBOW），兩者特性和適用情境並不相同。Jang et al. (2019) 指出，Skip-gram 架構在處理短篇、語境稀疏的文本分類上表現更佳。CBOW 則較適合長文本與語境完整的資料。綜上所述，本研究所採用的 Word2Vec 模型皆基於 Skip-gram 架構。主要考量到 K-12 口語材料多為句子短、語境分散，Skip-gram 將有助於提升語義向量於文本難度分級上的效能。本研究分別選中央研究院開發之 Word2Vec 模型，詞嵌入維度為 300，其餘參數為預設，訓練語料來自 Chinese Gigaword 與 ASBC 語料庫，詞彙量約 517,015 詞；另一為奧斯陸大學提供之 Word2Vec 模型，同樣採用 Skip-gram 架構，詞嵌入維度為 100，視窗大小為 10，其餘為預設，訓練語料來自 ChineseT CoNLL17 語料庫，詞彙量約

1,935,503 詞。經由不同斷詞工具與語言模型的組合，本研究最終產生四種特徵組合。表 2 所示分別為：（1）Jieba 斷詞搭配中央研究院 Word2Vec 後稱 J-AS、（2）Jieba 斷詞搭配奧斯陸大學 Word2Vec 後稱 J-OS、（3）CkipTagger 斷詞搭配中央研究院 Word2Vec 後稱 C-AS，以及（4）CkipTagger 斷詞搭配奧斯陸大學 Word2Vec 後稱 C-OS。每一組皆以詞嵌入平均（Mean Pooling）的方式生成文本語義向量，作為後續分級模型的特徵輸入。

最終，所有組合產生的文本語義向量，皆統一作為特徵輸入支持向量機。支持向量機採用 Scikit-Learn (Pedregosa et al., 2011)套件，核函數（Kernel）設定為徑向基函數（Radial Basis Function，RBF）其餘參數維持預設，並以五折交叉驗證（5-Fold Cross-Validation）(Arlot & Celisse, 2010)訓練 K-12 年級口語文本的可聽性模型。本研究藉由系統性比較各組特徵組合的分級效能，以探討斷詞策略及語言模型對於分級模型表現的影響。

年級	總數
K	715
1	2
2	14
3	43
4	31
5	49
6	20
7	209
8	222
9	233
10	460
11	597
12	270

表 1. 各年級影片總數

斷詞工具	語言模型	模型名稱
Jieba	中央研究院 Word2Vec	J-AS
	奧斯陸大學 Word2Vec	J-OS
CkipTagger	中央研究院 Word2Vec	C-AS
	奧斯陸大學 Word2Vec	C-OS

表 2. 斷詞工具與語言模型組合及對應模型名稱

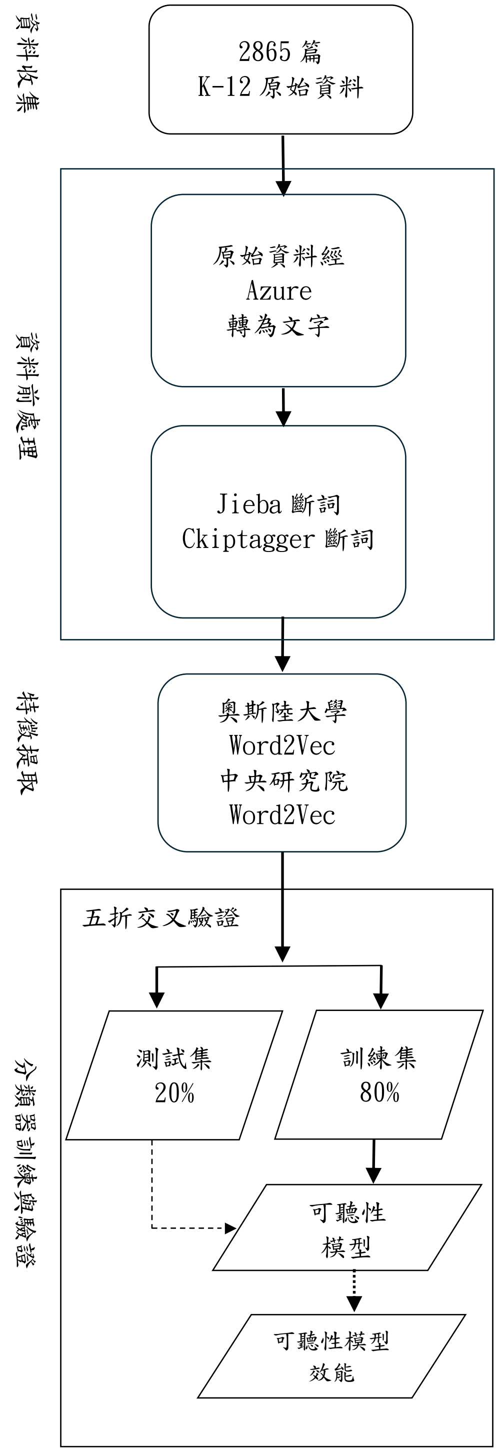


圖 1. 基於不同斷詞及 Word2Vec 可聽性模型訓練流程圖

4 研究結果

本節將分析不同斷詞工具與不同語言模型組合下，可聽性模型於 K-12 年級分級任務的效能表現。為檢驗模型效能的穩定性，所有效能指標均採用五折交叉驗證所得結果。評估

指標包括準確率（Accuracy）、鄰近準確率（Adjacent Accuracy）及混淆矩陣（Confusion Matrix）。其中，準確率衡量模型預測完全正確的比例；鄰近準確率則放寬標準，將預測年級與實際年級落差在前後一個年級內亦視為正確，有助於觀察模型分級錯誤的情形。混淆矩陣則能以直觀方式呈現模型在不同年級間的分級混淆狀況，幫助檢視哪些年級最易被誤判為其他年級，進而分析模型分級的困難點。

4.1 Jieba 斷詞工具與不同與語言模型之分級效能比較

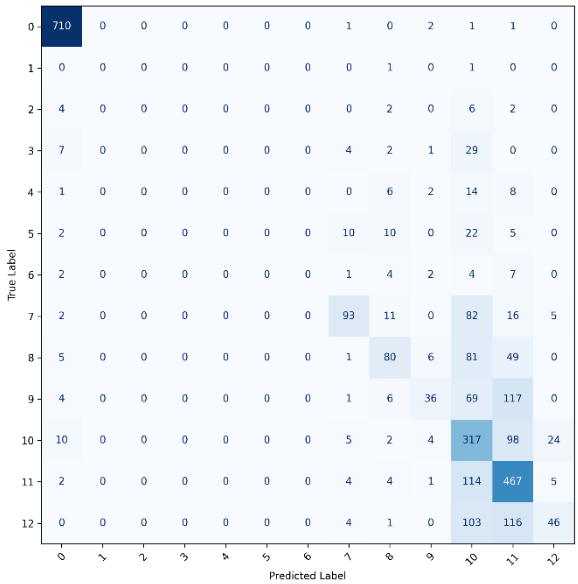
本小節比較了 Jieba 斷詞下， J-AS 與 J-OS 兩組模型於 K-12 年級分級任務的整體表現。表 3 中呈現的指標包括：準確率與鄰近準確率。這兩項指標可用以整體評估模型在 K-12 年級分級任務中的效能，亦能反映模型對於相近年級語料的實際預測能力。從數據結果來看，J-AS 的整體準確率與鄰近準確率分別為 61%和 76%，皆略高於 J-OS 的 59%與 74%。顯示出較高維度的詞嵌入有助於提升分級效能。不過，兩組模型整體表現相近，這可能與奧斯陸大學 Word2Vec 模型的詞彙量較高（約 1,935,503 詞），而中央研究院 Word2Vec 僅約 517,015 詞有關。較高的詞彙量可能使奧斯陸大學模型在詞嵌入維度較低的情況下，依然能維持一定的分級效能。整體來看，在 Jieba 斷詞下，詞嵌入維度與詞彙量皆可能對模型表現產生影響。

為進一步剖析模型分級，圖 2 展示了 J-AS 與 J-OS 兩組模型五折交叉驗證後的平均混淆矩陣。混淆矩陣可直觀呈現模型在各年級間的分級結果，縱軸為實際年級，橫軸為預測年級，對角線數值越高表示預測準確度越佳，對角線以外則代表誤判情形。藉由觀察混淆矩陣，不僅能了解模型在整體上的分級效能，也能發現模型容易混淆的年級區間。從圖 2 兩組模型的混淆矩陣結果可以發現，對於學齡前（K）年級的辨識表現最為突出，多數 K 年級語料均能正確分級。然而，國小 1 至 6 年級語料較常被誤分至 K 年級或高年級。顯示該區間語料分級錯誤可能被 K 年級和高年級樣本主導，此種誤判現象也可見於部分 K 年級與高年級資料。此外，自 7 年級起，模型預測結果顯著集中於對角線，10 至 12 年級的分級成效較為穩定，唯 9 及 12 年級仍有部分誤判。整體來看，混淆矩陣反映出樣本數較多或分布較集中的年級，其分級準確度較高，此外，低年級與 K 年級或高年級之間的誤判現象也較為明顯，顯示資料本身的分布特性會影響模型的分級效果。

模型	J-AS	J-OS
準確率	0.61	0.59
鄰近準確率	0.76	0.74

表 3. Jieba 斷詞下不同語言模型之分級效能比較

（a）



（b）

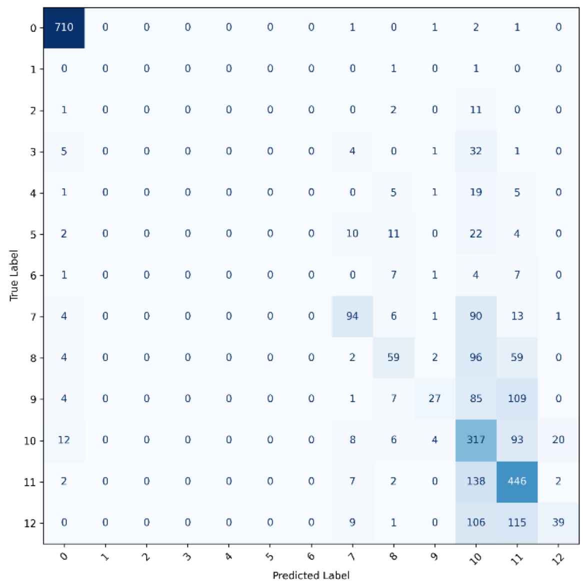


圖 2. 實驗 1 之各年級語料混淆矩陣 （a）J-AS 模型；（b）J-OS 模型

4.2 CkipTagger 斷詞工具與不同語言模型之分級效能比較

本小節比較了 CkipTagger 斷詞下，C-AS 與 C-OS 兩組模型於 K-12 年級分級任務的整體表現。在進一步比較 C-AS 與 C-OS 分級效能前，先說明表 4 中的 OOV（Out of Vocabulary）指標。OOV 指的是斷詞結果中，未能在 Word2Vec 詞彙表中找到向量的比例。OOV 比例越低，代表斷詞結果中的詞彙能夠被語言模型覆蓋。根據表 4 所示，CkipTagger 斷詞搭配其他 Word2Vec 模型時 OOV 比例都低於 Jieba 斷詞，顯示 CkipTagger 在中文口語材料的詞彙覆蓋表現較佳，能提升語義向量的完整度。

然而，根據表 5 兩組模型數據，C-AS 與 C-OS 有著更低的 OOV 比例，但準確率與鄰近準確率卻沒有提升，反而些微降低。此現象顯示，以本研究而言，降低 OOV 並未提升可聽性模型的效能。在現有的模型設計下，即使詞彙覆蓋率增加，模型對年級特徵的區辨力依然有限。因此，分級效能的提升或許還需結合更多語言特徵或更細緻的向量表示設計，方能發揮 CkipTagger 斷詞詞彙覆蓋率的優勢。此外，圖 3 二組混淆矩陣也顯示，低年級語料仍頻繁誤判為 K 年級或高年級，誤判形勢與 Jieba 斷詞結果高度相似，進一步說明提升詞彙覆蓋率對分級困難區間的改善有限。綜合來看，CkipTagger 斷詞能顯著降低 OOV 比例，但在目前模型設計下，對整體分級效能提升有限。因此，本研究建議未來可嘗試結合更進階的語言模型或分級器，以探討是否能進一步提升模型整體效能，以及加強對個別年級的預測能力。

模型	OOV 比例
J-AS	9.92%
J-OS	15.19%
C-AS	3.25%
C-OS	12.01%

表 4. 各組模型之平均 OOV 比例

模型	C-AS	C-OS
準確率	0.60	0.58
鄰近準確率	0.75	0.74

表 5. CkipTagger 斷詞下不同語言模型之分級效能比較

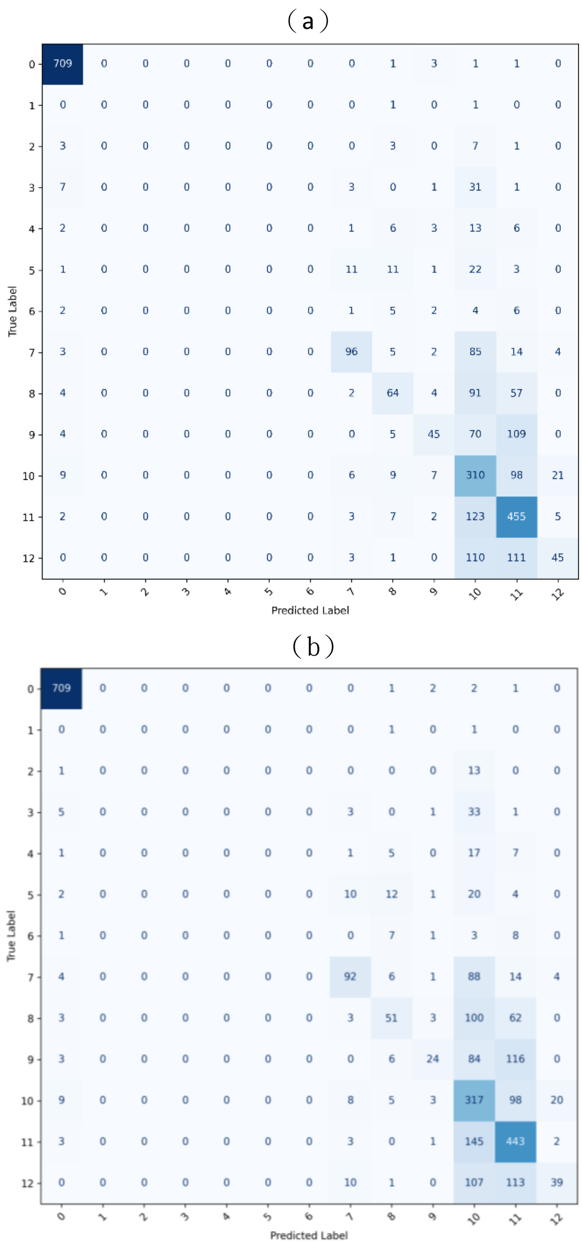


圖 3. 實驗 2 之各年級語料混淆矩陣 （a）C-AS 模型；（b）C-OS 模型

5 結論與未來發展

本研究針對中文 K-12 年級口語材料，系統性探討不同斷詞工具與語言模型組合下，口語可聽性自動分級模型之效能表現。主要研究內容包括：收集涵蓋學齡前至高中三年級的中文口語材料，並以 Jieba、CkipTagger 兩套中文斷詞工具及中央研究院與奧斯陸大學 Word2Vec 語言模型，組合成四組分級特徵進行效能比較。所有模型效能均採用五折交叉驗證進行評估，指標包括準確率、鄰近準確率以及混淆矩陣。實驗結果顯示，在所有組

合中，以 J-AS 表現最佳，整體準確率達 61%、鄰近準確率高達76%。同時，本研究發現詞嵌入維度較高的語言模型，在年級分級任務上表現普遍優於維度較低的模型。相較之下，CkipTagger 斷詞工具雖能有效降低 OOV 比例，但在現有設計下，並未顯著提升分級準確率。此外，混淆矩陣結果也揭示，低年級語料較容易被誤判為K年級或高年級，而高年級分級成效較為穩定。

在學術貢獻上，本研究彌補了中文口語可聽性模型領域中，針對斷詞工具與語言模型組合比較之研究空白，提供系統性實證結果，亦為未來中文可聽性模型帶來一定參考價值。然而，值得注意的是，現行大多國小至高中的教材內容多採用螺旋式設計（Spiral Curriculum），即同一主題會在不同年級反覆出現，但難度、深度與用詞豐富度逐步提升(Bruner, 2009)。例如，「水循環」這一主題在低年級的教材中，可能僅以簡單的圖畫介紹水的蒸發與降雨，而在高年級則會進一步說明蒸發、凝結、降水等科學原理，甚至討論水資源管理等議題。若分級模型僅依賴「水循環」這一關鍵詞，便可能將內容較淺顯的教材誤判為高年級，進而降低分級準確度。這種設計雖有助於學生循序漸進地建立知識體系，但當分級模型僅應用語言模型進行分級時，過度依賴主題關鍵詞可能導致分級不準確，容易將內容簡單的教材誤分為高年級。

回顧相關文獻，過去可聽性模型多結合語言特徵與語音特徵，本研究則以語言模型為主要特徵。未來建議可以進一步結合語言特徵與語音特徵，強化模型在部分年級分級上的區分能力。另外，本研究雖以較早期的語言模型 Word2Vec 為主，然而透過系統性比較不同斷詞工具與詞嵌入模型的組合效能，彌補現有口語可聽性模型分級領域的不足，未來亦可在此基礎上，導入如 BERT、Longformer 等語言模型，提升中文口語可聽性模型的分級效能。

6 致謝

本研究承國科會研究計畫「114-2628-H-011-002-MY3」、國立臺灣科技大學教育部高教深耕計畫特色領域技職賦能研究中心及國立臺灣師範大學教育部高教深耕計畫華語文科技中心補助，並感謝中央研究院詞庫小組提供語言資源，謹此致謝。

參考文獻

Alghamdi, E. A., Gruba, P., Masrai, A., & Velloso, E. (2023). The use of lexical complexity for assessing difficulty in instructional videos. https://hdl.handle.net/10125/73524

Alghamdi, E. A., Gruba, P., & Velloso, E. (2022). The relative contribution of language complexity to second language video lectures difficulty assessment. *The Modern Language Journal*, *106*(2), 393-410. https://doi.org/10.1111/modl.12773

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. https://doi.org/10.1214/09-SS054

Bahmani, R., & Farvardin, M. T. (2017). Effects of Different Text Difficulty Levels on EFL Learners' Foreign Language Reading Anxiety and Reading Comprehension. *Reading in a foreign language*, *29*(2), 185-202. https://files.eric.ed.gov/fulltext/EJ1157550.pdf

Bayona, M. G. A., Hines, A., Gilmartin, E., & Dhonnchadha, E. U. (2023). An Evaluation of the Use of Text-Based Comprehensibility Measures on Online Spoken Language Learning Materials. In 2023 34th Irish Signals and Systems Conference (ISSC), https://doi.org/10.1109/ISSC59246.2023.10162065

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*. https://doi.org/10.48550/arXiv.2004.05150

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Bruner, J. S. (2009). *The process of education*. Harvard university press. https://doi.org/10.4159/9780674028999

Cambria, J., & Guthrie, J. T. (2010). Motivating and engaging students in reading. *The NERA journal*, *46*(1), 16-29.

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. https://doi.org/10.1044/1092-4388(2006/023)

Chall, J. S. (1983). Stages of reading development. https://www.academia.edu/download/56874484/stages_of_reading_development.pdf

Chen, C.-Y., & Ma, W.-Y. (2018). Word embedding evaluation datasets and wikipedia title embedding for Chinese. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), https://aclanthology.org/L18-1132/

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, 37-54. https://www.jstor.org/stable/1473669

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, *26*(1), 19-26. https://www.jstor.org/stable/41383594

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. 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), https://doi.org/10.18653/v1/N19-1423

Ehri, L. C., & Wilce, L. S. (1985). Movement into reading: Is the first stage of printed word learning visual or phonetic? *Reading Research Quarterly*, 163-179. https://doi.org/10.2307/747753

Fang, I. E. (1966). The "Easy listening formula". *Journal of Broadcasting & Electronic Media*, *11*(1), 63-68. https://doi.org/10.1080/08838156609363529

Fares, M., Kutuzov, A., Oepen, S., & Velldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In Proceedings of the 21st nordic conference on computational linguistics, https://aclanthology.org/W17-0237/

Fariani, R. I., Junus, K., & Santoso, H. B. (2023). A systematic literature review on personalised learning in the higher education context. *Technology, Knowledge and Learning*, *28*(2), 449-476. https://doi.org/10.1007/s10758-022-09628-4

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221. https://doi.org/10.1037/h0057532

fxsjy. (2012). *Jieba*. Retrieved July from https://github.com/fxsjy/jieba

Gunning, R. (1952). The technique of clear writing. *(No Title)*. https://cir.nii.ac.jp/crid/197114938474081142 8

Harwood, K., & Cartier, F. (1952). On definition of listenability. *Southern Journal of Communication*, *18*(1), 20-23. https://doi.org/10.1080/10417945209371245

Hogan, T. P., Adlof, S. M., & Alonzo, C. N. (2014). On the importance of listening comprehension. *International journal of speech-language pathology*, *16*(3), 199-207. https://doi.org/10.3109/17549507.2014.904441

Jang, B., Kim, I., & Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, *14*(8), e0220976. https://doi.org/10.1371/journal.pone.0220976

Kayam, O. (2018). The readability and simplicity of Donald Trump's language. *Political Studies Review*, *16*(1), 73-88. https://doi.org/10.1177/1478929917706844

Kincaid, J. P., Fishburne Jr, 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. https://doi.org/10.21236/ADA006655

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., & Aluísio, S. M. (2024). NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, *58*(1), 73-110. https://doi.org/10.1007/s10579-023-09693-w

Li P-H, M. W.-Y. (2019). *CkipTagger*. Retrieved July from https://github.com/ckiplab/ckiptagger

Liu, Y., Ji, M., Lin, S. S., Zhao, M., & Lyv, Z. (2021). Combining readability formulas and machine learning for reader-oriented evaluation of online health resources. *IEEE Access*, *9*, 67610-67619. https://doi.org/10.1109/ACCESS.2021.3077073

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In Proceedings of the Annual Meeting of the Cognitive Science Society, https://escholarship.org/content/qt7d8631cr/qt7d8631cr.pdf

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, *12*(8), 639-646. https://www.jstor.org/stable/40011226

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In Proceedings of the 20th SIGNLL conference on computational natural language learning, https://aclanthology.org/K16-1006.pdf

Microsoft. (2025). Azure AI 語音. Retrieved 8 from https://azure.microsoft.com/zh-tw/products/ai-services/ai-speech

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://doi.org/10.48550/arXiv.1301.3781

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier. https://books.google.com/books?hl=zh-TW&lr=&id=mn2jBQAAQBAJ&oi=fnd&pg=PP1&dq=Probabilistic+reasoning+in+intelligent+systems:&ots=4tFU2E8M90&sig=ww1O97ittFH8heGbsDUlmuGFRRY

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830. https://doi.org/10.5555/1953048.2078195

Peppa-Pig-Chinese-Official. (n.d.). 小猪佩奇中文官方 Peppa Pig Youtube 頻道. Youtube. Retrieved 7 from https://www.youtube.com/@PeppaPigChineseOfficial

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer speech & language*, *23*(1), 89-106. https://doi.org/10.1016/j.csl.2008.04.003

PTSKIDS. (n.d.). 小公視 Youtube 頻道. Youtube. Retrieved 7 from https://www.youtube.com/@PTSKIDS

Rogers, J. R. (1962). A formula for predicting the comprehension level of material to be presented orally. *The journal of educational research*, *56*(4), 218-220. https://doi.org/10.1080/00220671.1962.10882926

Uçar, S.-Ş., Aldabe, I., Aranberri, N., & Arruarte, A. (2024). Exploring automatic readability assessment for science documents within a multilingual educational context. *International Journal of Artificial Intelligence in Education*, *34*(4), 1417-1459. https://doi.org/10.1007/s40593-024-00393-2

XIAOXINGXING-樂樂 TV. (n.d.). 樂樂 TV Youtube 頻道. Youtube. Retrieved 7 from https://www.youtube.com/@XIAOXINGXING-%E6%A8%82%E6%A8%82TV

Yin, Z., & Shen, Y. (2018). On the dimensionality of word embedding. *Advances in neural information processing systems*, *31*. https://papers.neurips.cc/paper/2018/file/b534ba68236ba543ae44b22bd110a1d6-Paper.pdf

Yoon, S.-Y., Cho, Y., & Napolitano, D. (2016). Spoken text difficulty estimation using linguistic features. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, https://doi.org/10.18653/v1/W16-0531

Zhang, C. (2024). RCS-CSLT: a language feature-driven readability classification system for international Chinese education. *Computer Assisted Language Learning*, 1-27. https://doi.org/doi.org/10.1080/09588221.2024.2430723

北鼻故事屋 YouTube 頻道. (n.d.). Retrieved 8 from https://www.youtube.com/channel/UCHNtUji8P8yz645YCgs8GUA

教育部因材網. (2020). Retrieved 0804 from https://adl.edu.tw/HomePage/home/

臺北酷課雲 Taipei CooC-Cloud. (n.d.). Retrieved 0804 from https://cooc.tp.edu.tw/

Cubicpower Agentic Mixture of Experts (AMoE) Framework for Fine-Tuning NLP Tasks Without GPUs

Chao-Yih Hsia(夏肇毅)
Smart Academy
CubicPower Smart Center
Taipei,Taiwan
chaoyihhsia@gmail.com

Abstract

The rise of Green AI emphasizes minimizing the environmental footprint of AI systems. This paper explores a no-GPU agentic architecture for fine-tuning NLP tasks. It presents our initial experiments applying these no-GPU algorithms in pretraining and fine-tuning tasks on our CubicPower agentic mixture of experts (AMoE) framework, with the aim of contributing to more sustainable AI development. In contrast to the training procedures of neural networks, which consume significant power, the AMoE framework's primary contribution toward power savings is that it requires no training process. We explore non-neural-network methods for solving NLP tasks and employ similarity measures to match predefined patterns for use in a RAG database.

Keywords: Green AI, MoE, RAG, CubicPower, AMoE.

1 Introduction

In recent years, many countries have set a 2050 net-zero emissions goal. Energy conservation has become a top priority across all industries. However, AI neural network algorithms, such as the Bitcoin Proof-of-Work (PoW) algorithm, rely heavily on GPUs or other custom-designed accelerators. These machine learning training processes, using the gradient descent method, can take weeks or months to run on large numbers of high-power-consuming GPUs. Therefore, many solutions have been developed to save energy (Verdecchia et al., 2023). However, we believe that a no-GPU Green AI algorithm could be a new and effective direction (Hsia, 2022), since it eliminates the primary source of power consumption.

Traditional text mining algorithms use parameters to measure word properties, such as TF-IDF and similarity. TF-IDF measures the importance of a word, while similarity measures the distance between words. These algorithms are not neural networks and, of course, do not involve any gradient descent training process. We have developed algorithms based on text similarity to select the most similar text from the pattern pool.

This paper presents our initial experiments applying such no-GPU algorithms in pretraining and fine-tuning tasks on our CubicPower agentic mixture of experts (AMoE) framework, aiming to contribute toward more sustainable AI development.

While MoE and RAG approaches have improved efficiency, most still rely on GPU computation. We propose a GPU-free AMoE framework using similarity-based retrieval to fine-tune NLP tasks.

The main contributions of this paper are as follows:

1. Exploration of non-neural-network methods for solving NLP tasks.
2. Elimination of the training process in the AMoE framework to save power.
3. Use of similarity measures to match predefined patterns for retrieval in a RAG database.

2 Related Work

Early dialogue systems evolved from rule-based methods, such as ELIZA (Weizenbaum, 1966), which applied pattern-matching rules to simulate human-like responses. This approach laid the foundation for later systems, such as GUS (Bobrow, 1977), which introduced a frame-based architecture. In GUS, dialogues were organized into structured templates containing slots, enabling simple task-oriented conversation handling.

Modern systems have shifted toward neural architectures. The sequence-to-sequence (seq2seq) model, originally designed for machine translation (Sutskever et al., 2014; Bahdanau et al., 2015), was later adapted for chatbot design. It uses an encoder-decoder structure and autoregressive generation. These models are typically powered by GPU-intensive training and inference pipelines.

To reduce computation costs, retrieval-based systems have re-emerged, using similarity metrics (e.g., cosine similarity) to find the most relevant response from a pattern database. This is often more power-efficient than generation-based models. Retrieval-Augmented Generation (RAG) combines neural language models with external information retrieval, offering enhanced relevance and scalability (Gao et al., 2023).

Similarity search plays a crucial role in these systems. Johnson et al. (2019) proposed a billion-scale similarity search framework using GPUs, while Han et al. (2023) surveyed vector databases and their indexing strategies. In contrast, Hsia (2022) developed a GPU-free similarity-based system, forming the basis of the CubicPower knowledge base, which enables fast and structured retrieval.

Another important concept for reducing computation is the Mixture of Experts (MoE). MoE architectures achieve scalability by activating only a small subset of the model's parameters for each input, allowing for high model capacity without proportional increases in computation. Shazeer et al. (2017) demonstrated this with the Sparsely-Gated MoE, where only a few experts are selected per example, reducing computational cost while preserving performance.

The rise of Green AI (Verdecchia et al., 2023) emphasizes minimizing the environmental footprint of AI systems. Techniques that reduce power consumption, including rule-based reasoning, task-specific similarity retrieval, and agent-level model decomposition, align with this goal. This paper explores a no-GPU agentic architecture for fine-tuning NLP tasks.

3 Methodology

In this paper, we develop the entire AMoE framework based on the CubicPower Data Processing Engine for similarity computation, following the description in Hsia (2022). The framework was implemented in C# .NET.

3.1 Agentic Architecture

We define AI agents as modular components, each responsible for a specific NLP fine-tuning task, such as question answering (QA), reading comprehension (RC), or chatbot dialogue state tracking. Each agent maintains a local dataset and operates independently, processing only the inputs relevant to its task domain. This follows a Mixture of Experts (MoE) model design but is implemented without neural networks.

3.2 Retrieval-Augmented Module

Figure 1. shows the design of our AMoE framework to perform the retrieval-augmented generation (RAG) function.

Each agent is equipped with a sentence-level retrieval mechanism. It consists of a vector database which stores sequence to sequence (seq2seq) pair records such as question-answers.

Given an input, the agent generates a corresponding sentence vector and compares it against stored records by dot-product to compute their similarities. Then the system finds the record i with the highest similarity. Extracting the second part of the seq2seq pair, we can find the answer to the question. By leveraging these structures, we can operate the retrieval-augmented generation (RAG) process effectively.

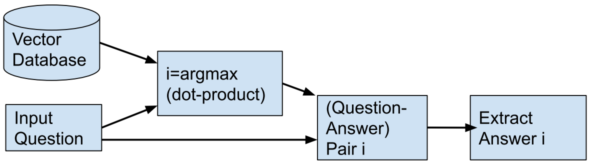


Figure 1. Our approach

3.3 Dataset and Procedures for Fine-Tuning Tasks

All datasets are stored in plain text format to ensure efficient loading and access by intelligent agents. This format facilitates rapid retrieval, parsing, and integration into downstream tasks such as question answering, multiple-choice tasks, and reading comprehension.

3.3.1 Question Answering (QA) Task:

The dataset for the QA task consists of question–answer pairs, as shown in Figure 2. We have collected sets of question–answer pairs. To perform the QA task, we need to analyze the QA training dataset to construct the overall word distribution. First, we sample the QA training dataset to construct the overall QA word distribution:

$$D=\text{Sample}(\text{QA training set}) = M_{W\text{->}P} \quad (1)$$

Here D is the distribution of the current QA words. This distribution is used to map each word to a sentence. The output is a word-to-paragraph map $M_{W\text{->}P}$. It is used to compute the most similar paragraphs from a group of words in a question.

Following the description in Hsia (2022), we can implement a similarity-based system using the distribution D to find paragraphs from a word. Each paragraph is a question–answer pair.

We then build a paragraph-based RAG module RAG_D to select answers from RAG_D for the questions.

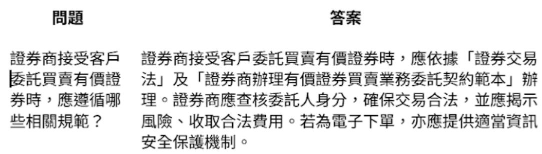
Denote $RAG_D()$ as a RAG module based on the distribution D. Once we feed a question into this module, the output paragraph from this module for a question becomes:

$$\text{paragraph}_{RAGD} = RAG_D(\text{question}) \quad (2)$$

We can therefore obtain the answer to the question as a QA RAG output answer:

$$\text{answer} = \text{answerOf}(\text{paragraph}_{RAGD}) \quad (3)$$

The answerOf() function in (3) returns the answer from a paragraph containing a question–answer pair.



問題	答案
證券商接受客戶委託買賣有價證券時，應遵循哪些相關規範？	證券商接受客戶委託買賣有價證券時，應依據「證券交易法」及「證券商辦理有價證券買賣業務委託契約範本」辦理。證券商應查核委託人身分，確保交易合法，並應揭示風險、收取合法費用。若為電子下單，亦應提供適當資訊安全保護機制。

Figure 2. Dataset for the question answering task

3.3.2 Multiple Choice (MC) Task:

The dataset consists of question–option–answer triples, where each record contains a question, options A–D, and the correct answer, as depicted in Figure 3.

Each multiple-choice question can be reformulated into four independent True or False questions, allowing the system to evaluate each option separately.

Alternatively, the task can be approached as a QA problem by checking the answer to existing questions in the training set.

For unseen questions, we must learn the question–answer relationships from the training set and select the option whose relationship most closely matches the learned patterns.

問題	選項1	選項2	選項3	選項4	答案
企業的主要目的是？	提供就業機會	追求利潤	創造藝術	進行研究	追求利潤

Figure 3: Dataset for the multiple choice task

3.3.3 Reading Comprehension (RC) Task:

Similar to the QA task, we need to analyze the word distribution for the RC task. However, the source of the word distribution is not the training set; it comes from each RC question. Therefore, we must resample the RC question each time to reconstruct the RC word distribution for that question.

In order to answer a question in the RC task, we first resample the RC document i in the test dataset to extract the word distribution D_i of the RC question i.

$$D_i = \text{Resample(RC document i)} = M_{i\ W\text{->}P} \quad (4)$$

Here D_i is the word distribution of the current RC question i. This distribution is used to map each word to a paragraph.

Following the same method as QA, we can build a paragraph-based RAG module RAG_{Di} to select answers for the questions.

Here we denote $RAG_{Di}()$ as a RAG module based on the distribution D_i. The output paragraph for a question becomes:

$$\text{paragraph}_{RAGDi} = RAG_{Di}(\text{question}) \quad (5)$$

We can therefore obtain the answer to the question from the RC RAG output:

$$\text{answer} = \text{answerOf}(\text{paragraphs}_{RAGDi}) \quad (6)$$

The answerOf() function in (6) returns the answer from a paragraph containing a question–answer pair.

閱讀內容	問題1	答案1	問題2	答案2	問題3	答案3
管理是指透過規劃、組織、領導與控制等程序，有效且效率地達成組織目標的過程。有效性代表能完成正確的工作，效率則是指以最少的資源完成任務。現代管理學者強調「以人為本」與「持續改善」的概念，認為管理不僅是資源配置，更涉及激勵員工與創造學習文化。	管理的四大功能有哪些？	規劃、組織、領導與控制	什麼是管理中的「效率」？	以最少資源完成任務	現代管理強調什麼價值觀？	以人為本與持續改善

Figure 4. Dataset for the reading comprehension task

3.3.4 Chatbot (CB) Task:

The chatbot dataset consists of paired utterances, each representing a conversational turn, as illustrated in Figure 5. The task involves predicting the next appropriate response based on the current user input.

聽到	回答
早安，今天的天氣如何？	早安！今天陽光明媚，氣溫約25度，非常適合外出。

Figure 5. Dataset for the chatbot task

3.4 Power-Efficient Design

In contrast to the training procedures of neural networks, which consume significant power, the main contribution of the AMoE framework to power saving is that it requires no training process.

Additionally, the CubicPower AMoE framework consists of many agents. Each agent stores only a small portion of data relevant to its task. This follows the Mixture of Experts (MoE) method (Lepikhin et al., 2020; Fedus et al., 2022). In our system, the experts are agents. Therefore, only a small amount of power is consumed at any given time. Furthermore, we can split the data by language, geographical location, and type, assigning each subset to a different agent. The system decides which agent should handle the input based on the content of the prompt.

4 Experiments

The experiment in this study relies on a similarity metric. Similarity is measured as the proportion of words in the correct answer that also appear in the predicted answer.

This measure is conceptually similar to BLEU-1 (Bilingual Evaluation Understudy), which assesses word overlap between reference and generated text.

4.1 Experimental Setup

All experiments were conducted on a standard CPU-based machine without GPU acceleration. The framework was implemented using C# .NET and utilized the CubicPower Data Processing Engine's classical text processing libraries for cosine similarity computation.

Each task-specific agent was evaluated independently using a dedicated dataset, split into training and testing subsets. The training set served as the retrieval base for the test queries.

4.2 Datasets

We prepared different datasets for fine-tuning tasks. We used small private datasets collected by CubicPower. Each dataset contains several hundred records.

For the Question Answering task, the dataset consists of a question and an answer field (see Figure 2). When a QA agent receives a QA request with a question, it searches the question field of the database and returns the most similar QA record.

For the Multiple Choice task, our dataset was prepared as shown in Figure 3. For each question, there are four options. The final field contains the answer to the question. Each question is essentially a combination of four true or false questions. By testing each of the four true or false questions, only one of them will be true.

The Reading Comprehension task first provides a document and then asks a series of questions based on that document.

We aim to answer the questions using only the material provided in the document; therefore, we need to build a word space derived from this document. Figure 4 shows a sample of the RC dataset.

Table 1 lists the sizes of the training and test sets for all four fine-tuning tasks used in our experiments.

Task	**Training Set Size**	**Test Set Size**
QA	749	371
MC	440	181
RC	—	619
CB	1121	389

Table 1: tasks train/test Dataset Size

5 Results

5.1 Fine-tuning Tasks Test:

We loaded the training dataset for the QA task into our database and then used it to verify the search results. Figure 6a shows a screenshot of the verification results on the training set. We can see that the top-1 accuracy is 0.847, and the similarity between the question and the returned answer is 0.983.

Then, we used the test dataset to query the training database. Figure 6b shows a screenshot of the test results. The results are near zero since there should be no overlap between the training and the test datasets. The nonzero result indicates that some data leakage exists between the two datasets.

Figures 7 to 9 show the remaining test result screenshots for the MC, RC, and CB tasks. Table 2 summarizes their test results.

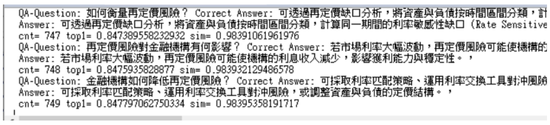


Figure 6a. Figure 6a. QA Train Verification Result

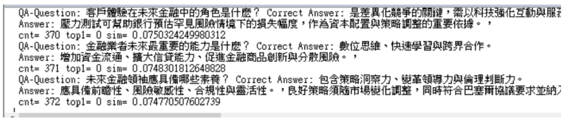


Figure 6b. QA Test Result

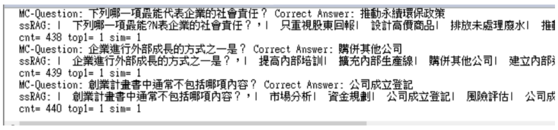


Figure 7a. MC Train Verification Result

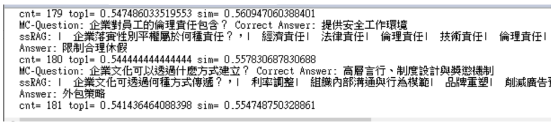


Figure 7b. MC Test Result

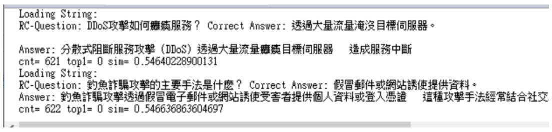


Figure 8. RC RAG Test Result

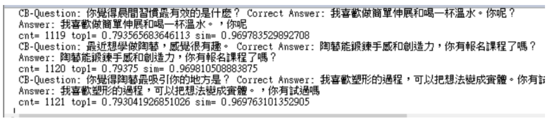


Figure 9a. CB Train Verification Result

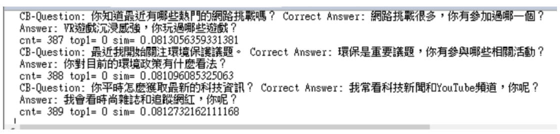


Figure 9b. CB Test Result

Table 2. Task Training/Test Similarity

Task	Similarity (Train)	Similarity (Test)
QA	0.983	0.074
MC	1	0.554
RC		0.546
CB	0.969	0.081

5.2 Exams Test:

We evaluated the performance of our AMoE system using three datasets. The first dataset includes the Taiwan government employee entrance tests and the Financial Institution Certification. The second dataset contains the Taiwan Government Professional Certifications. The third dataset is the Taiwan Massive Multitask Language Understanding Plus (TMMLU+) dataset (Tam et al., 2024).

Figure 10 shows screenshots of the test results, and Table 3 summarizes these results. The first test includes 33,608 training records and achieves an accuracy of 0.354. The second test contains 20,807 training records, achieving an accuracy of 0.283. The third test has 21,120 records and achieves a test accuracy of 0.289.

Table 3. Exam Test Results

	Train set	Data Set	MC Task Accuracy
Financial Institution Certifications / government employee entry test.	33,608	26,985	0.354
Government Professional Certifications	20,807	2,069	0.283
TMMLU+	21,120	2,225	0.289

5.3 Benchmarking Test:

To compare the performance with other Traditional Chinese LLM models, we tested the TMML+ benchmark dataset using zero-shot and 5-shot settings.

Table 4 presents the TMML+ benchmark results for different LLM models reported by Tam et al. (2024). The results show that the zero-shot average accuracy of Breeze-7B-Instruct-v1.0 is 36.1%, which is higher than our 25.1%. However, the other two models, Taiwan-LLaMa-13B and Taiwan-LLaMa-7B, achieved accuracies of 21.3% and 15.6%, respectively. The performance of our AMoE framework in the Traditional Chinese TMMLU+ test ranks second among the compared models.

Table 4. Comparative Results on TMMLU+: (*from Tam et al., 2024)

LLM Models	Zero-shot accuracy (%)	5-shot accuracy (%)
*Breeze-7B-Instruct-v1.0	36.1	28.6
CubicPower AMoE	**25.1**	**25.7**
*Taiwan-LLaMa-13B	21.3	22.3
*Taiwan-LLaMa-7B	15.6	5.1

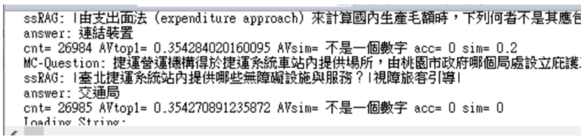


Figure 10a. Financial Institution Certifications / Government Employee Entry Tests

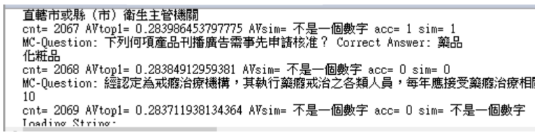


Figure 10b. Government Professional Certifications

Figure 10c. TMMLU+ Test Results

5.4 Discussion

The results indicate that the AMoE framework performs poorly on unseen data in the QA and CB tasks. One possible improvement is to expand the scope of the training dataset.

Additionally, the MC task accuracies in the Government Professional Certifications and TMMLU+ datasets are around 0.28, which is only slightly above random guessing. Although we rank second in the TMMLU+ Traditional Chinese test, there is still considerable room for improvement.

These challenging tests require extensive reasoning before an answer can be generated. As a result, it is difficult to apply a simple QA-style predefined answer list to solve them.

To address this, our next step will be to develop a reasoning agent that applies the chain-of-thought (CoT) method to complex problems.

6. Conclusion

The rise of Green AI emphasizes minimizing the environmental footprint of AI systems. Techniques that reduce power consumption, including rule-based reasoning, task-specific similarity retrieval, and agent-level model decomposition, align with this goal. Traditional text mining algorithms use parameters to measure word properties, such as similarity. We propose a GPU-free AMoE framework using similarity-based retrieval to fine-tune NLP tasks.

This paper explores a no-GPU agentic architecture for fine-tuning NLP tasks. It presents our initial experiments applying these no-GPU algorithms in pretraining and fine-tuning tasks on our CubicPower agentic mixture of experts (AMoE) framework, with the aim of contributing to more sustainable AI development. In contrast to the training procedures of neural networks, which consume significant power, the AMoE framework's primary contribution to power savings is that it requires no training process. We have developed basic functionalities, but there is still room for improvement. To address this, the next step of our research will be to develop a reasoning agent using the chain-of-thought (CoT) method for complex problems.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. https://arxiv.org/abs/1706.03762

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137–1155. https://doi.org/10.1162/153244303322533223

Bobrow, D. G. (1977). GUS, a frame-driven dialog system. *Artificial Intelligence*, 8(2), 155–173. https://doi.org/10.1016/0004-3702(77)90018-2

Cheng, H., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Shah, H. (2016). Wide & deep learning for recommender systems. *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems* (DLRS 2016). https://doi.org/10.48550/arxiv.1606.07792

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2101.03961

Gao, Y., Zhao, Y., Zhang, Y., Liu, Z., & Ding, G. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint*. https://simg.baai.ac.cn

Han, Y., Liu, C., & Wang, P. (2023). A comprehensive survey on vector database: Storage and retrieval technique, challenge. *arXiv preprint*. https://arxiv.org/abs/2310.11703

Hsia, C.-Y. (2022). Design of CubicPower real-time topic writing knowledge base system based on similarity (以相似度為基礎之CubicPower即時主題寫作知識庫系統設計) [Conference presentation]. *TANET 2022 Taiwan Internet Seminar*, Taiwan. https://drive.google.com/open?id=13PQnzzDIHSEFTfeX4NYWAMydlP0MwNh8

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 5(2), 1–11. https://doi.org/10.1109/TBDATA.2019.2902270

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., ... & Chen, Z. (2020). GShard: Scaling giant models with conditional computation. *arXiv preprint*. https://arxiv.org/abs/2006.16668

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*. http://arxiv.org/abs/1301.3781

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint*. https://arxiv.org/abs/1701.06538

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27, 3104–3112. http://cs224d.stanford.edu/papers/seq2seq.pdf

Tam, Z.-R., Pai, Y.-T., Lee, Y.-W., Chen, J.-D., Chu, W.-M., Cheng, S., & Shuai, H.-H. (2024). TMMLU+: An improved Traditional Chinese evaluation suite for foundation models. arXiv.

https://doi.org/10.48550/arXiv.2403.01858

Verdecchia, R., Sallou, J., & Cruz, L. (2023). A systematic review of Green AI. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1507

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. https://doi.org/10.1145/365153.365168

Design and Evaluation of a Courtroom Examination AI Simulation System with Behavioral Fidelity
具行為擬真度之法庭詰問 AI 模擬系統設計與評估

Liao, Hsien-Jyh
Independent Researcher
hjliao123@gmail.com

摘要

本研究提出一套以「行為擬真度 (Behavioral Fidelity)」為核心的法庭詰問 AI 模擬系統，並將語音互動設計納入以提升沉浸感與臨場再現；惟為確保評估之標準化與可重現性，本輪前導評估採文字筆錄。系統整合語用—心理規則與臺灣刑案卷宗，以模擬證人在詰問壓力下的語言行為；效能評估採經優化之「專家圖靈測試」框架，涵蓋四個面向：專業準確度、情境適應、人味化與邏輯一致性。作為前導研究（pilot），結果顯示：在相同題庫與知識依據下，客製化 GPT 相較於 GPT-Vanilla 於「情境適應」與「人味化」呈現較高評分；同一框架應用於另一主流模型（Gemini 2.5 Flash）時亦達可比水準，惟在本樣本規模下差異未形成一致結論。整體而言，研究提供「行為擬真度」作為評估指標之初步證據，並顯示生成式 AI 於法律訓練場域具可規模化應用之潛力；語音條件下之評估及多案件、多角色擴充留待後續研究。

Abstract

We present a courtroom cross-examination AI simulation system centered on Behavioral Fidelity, with speech interaction included as a design feature to enhance immersion. For standardization and reproducibility, the present pilot evaluation uses transcripts. The system integrates pragmatic–psychological rules with Taiwanese criminal case files to simulate witness behavior under cross-examination pressure. Using an optimized Expert Turing Test framework with four dimensions—professional accuracy, situational adaptability, human-likeness, and logical consistency—we conduct a pilot study. Under identical prompts and knowledge sources, the customized GPT condition received higher ratings than GPT-Vanilla on adaptability and human-likeness. Applying the same framework to another mainstream model (Gemini 2.5 Flash) yielded comparable performance, while differences remain inconclusive at this sample size. Overall, the results provide preliminary evidence that Behavioral Fidelity is a feasible evaluation target and indicate the scalability of generative AI for legal training; speech-condition evaluation and multi-case, multi-role extensions are left for future work.

關鍵字：法庭詰問、行為擬真度、法庭語言學、人工智慧模擬

Keywords: courtroom examination, behavioral fidelity, forensic linguistics, AI simulation

1 引言（Introduction）

在法律實務訓練中，交互詰問是塑造訴訟攻防與臨場決策的核心能力；然而，交互詰問難以實作練習，而傳統模擬法庭也因人力與環境限制，難以長期、規模化實施，但近年生成式人工智慧（Generative AI, GenAI）提供了新的路徑：除可低成本產生多樣情境，亦能扮演特定專業角色，支援反覆演練與即時回饋的教學流程。基此，本研究結合 GenAI、法庭真實語境及語音互動，開發一套可透過口語進行沉浸式詰問演練的法庭詰問 AI 模擬系統，並以「行為擬真度」(Behavioral Fidelity) 做為本系統的核心設計與評估理念。

然而，如何評估此類系統的有效性，仍是關鍵挑戰。現行主流評估（如 MMLU：

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 20–28
November 20-22, 2025 ©2025 Association for Computational Linguistics

Hendrycks et al., 2021；HELM：Liang et al., 2022）與傳統文本品質指標（ROUGE：Lin, 2004；BLEU：Papineni et al., 2002）多著眼於內容正確性與流暢度。相較之下，交互詰問的關鍵在於語用策略、權力互動與壓力下的臨場反應（Cotterill, 2003；Coulthard & Johnson, 2017；Gudjonsson, 2003；Slater & Sanchez-Vives, 2016）。若無法重現這些行為特徵，模擬的教學價值將受限。因此，本研究將焦點前移：從「答得對」轉向「像真人一樣在法庭上作答」。

為了在可重現前提下檢驗系統效能，本輪前導評估（pilot）採文字筆錄作為評估材料（語音互動僅為系統功能，非本輪評估對象），並設計「專家圖靈測試」式的人評框架，涵蓋四個維度：專業準確度、情境適應、人味化、邏輯一致性。我們以去識別化之臺灣刑案卷宗為知識依據，結合語用—心理規則與五大題型分類，生成在不同詰問階段下可控的證人行為。

本研究的主要貢獻如下：

(1) 提出並操作化「行為擬真度」作為評估核心：補足既有評估偏重內容正確性的不足，將臨場語用與行為一致性納入量化與質化的綜合指標。
(2) 建構一可重現的跨域系統架構：整合語用—心理規則、五大問題分類機制與本土卷宗知識庫，以支援高擬真度的口語互動演練。
(3) 以專家圖靈測試框架完成前導實證：在相同題庫與知識依據下，比較客製化 GPT 與 GPT-Vanilla 之差異，並提供與另一主流模型的敘述性對照；結果呈現估計導向報告（含效果方向與不確定性），為後續語音條件、跨案件與多角色之擴充研究奠定基準。

2 文獻回顧

本研究旨在建構一套法庭詰問 AI 模擬系統，其核心設計理念為行為擬真度（Behavioral Fidelity）。此目標連結法庭語言學、法庭心理學與人機互動／AI 角色模擬三大脈絡，焦點從「知識答對」前移至「在壓力下像真人一樣作答」，以支援可重現且具臨場感的專業訓練。

2.1 法庭詰問的挑戰：策略、權力與心理壓力

法庭詰問並非單純事實問答，而是結合提問策略、話語權力與心理壓力的互動博弈（Mauet, 2017；Cotterill, 2003）。律師得以透過封閉式／誘導式提問、重述與打斷等手法，塑造敘事框架與焦點（Cotterill, 2003；Coulthard & Johnson, 2017）。證人在高壓情境下常出現遲疑、模糊化、記憶偏移等自然反應（Gudjonsson, 2003），這些語用與心理特徵正是法律訓練的目標能力之一。據此，一個擬真的 AI 證人，必須能在題型與情境變化間，展現權變性的語用行為，而非僅生成流暢文本。

2.2 從角色扮演到「行為擬真」：AI 模擬的設計要件

專業訓練中常見的「標準化角色」（Barrows, 1993）提供了可重複的演練基線；但在 AI 應用場域，僅具語言流暢不足以令人信服。過度完美與缺乏內在情感邏輯的回應會引發恐怖谷效應（Mori, 2012），降低沉浸與信任。人機互動研究指出，可被感知的行為真實性能顯著提升臨場感與可信度（Slater & Sanchez-Vives, 2016）。因此，系統設計應將「恰當的不完美」納入（如遲疑詞、修辭回溯、語速與停頓），並透過可控的人設／語氣設定與程序性規則，使行為在不同詰問階段具一致的內在因果與邏輯邊界。

2.3 人類評估的必要性

當 AI 生成回應的目標是「行為擬真」而非僅「內容正確」，傳統自動指標（如 ROUGE、BLEU）對「是否像真人」的辨識力有限。對話系統評估文獻強調，人類評估仍是風格、自然度、權變性與一致性等主觀構念的黃金標準（Celikyilmaz et al., 2020；Deriu et al., 2021）。本研究據此改造了圖靈測試思路，引入專家圖靈測試框架，並將評估構面收斂為四類：專業準確度、情境適應、人味化、邏輯一致性。此處的人評並非臨時主觀打分，而是有理論錨定且可操作化的評分規準（見表 1）。

表 1：評估構面之理論錨定與可操作化指標

評估構面	理論錨定	主要觀察點（可操作化）
專業準確度	忠實性／事實一致性	僅援引卷宗或可驗證事實，避免範圍主張
情境適應	權變性／對話依存	（封閉／誘導／推論）題型對應；（誘導、假設不明時）拒答行為觸發；關鍵訊息之澄清與優先順序。
人味化	自然度／風格	遲疑詞率與語助詞使用；停頓樣式與句長變化；受控的不完美（冗詞、回溯／自我修正）。
邏輯一致性	前後一致／無矛盾	跨回合自我一致；與已述事實不矛盾 （可選）NLI 旗標數以輔助標記矛盾。

註：NLI（Natural Language Inference）可作為半自動化輔助，用於標示回答間的矛盾候-選，供評估者複核。

2.4 近期相關工作：AI 驅動之法庭模擬與辯論系統

近期開始出現若干以 GenAI 支援法學教育或辯論策略的系統，但研究目標與評估焦點與本研究不同：

(1) 教學框架類：如 Moot MentorAI（Serra, 2024）著重教學部署與文字論述回饋，評估主軸多在文本品質與學習成效，較少觸及「擬真行為」的量化。

(2) 多代理對抗類：如 AgentCourt（Chen et al., 2025）專注於對抗式演化以提升辯論策略，目標是「勝率／推理能力」而非「擬真行為」。

(3) 原型公告類：如 Stanford CodeX × Three Crowns 公告之訓練原型，尚未公開可重現之系統設定與評估數據。

整體而言，現有研究或偏向教學流程或文字技巧，或僅瞄準策略最適化，能同時兼顧行為擬真與語音互動設計，並提供透明可重現的評估框架者仍少見。本研究即補上此一縫隙：在單一 AI 證人的詰問場景中，以行為擬真為首要目標，提出可操作的人評標準與可重現的系統實作。

2.5 小結

綜合上述脈絡，既有研究或著重教學流程與文字技巧，或以多代理對抗提升辯論策略，較少將「行為擬真」作為首要評估目標，亦缺乏可重現之框架以量化法庭情境下的語用行為。基於此缺口，本研究提出以行為擬真度為核心的法庭詰問 AI 模擬系統，結合本土卷宗 RAG、語用—心理規則與角色／語氣設定，並採用專家圖靈測試式人評作為前導評估。需強調的是：語音互動為系統功能，本輪為確保標準化而採文字筆錄進行評估；外部效度（多案件、多角色）留待後續擴充。以下（第 3 章）據此說明系統架構與實作細節。

3 系統架構與實作

3.1 整體流程與技術核心

本系統以大型語言模型（LLM）為核心，整合語用—心理判斷規則、五類問題分類機制與本土刑事卷宗知識庫，形成一跨學科的智能模擬系統。為提升臨場感，系統整合了語音生成（TTS）與辨識（ASR）模組，讓使用者能以口語和 AI 證人互動。

整體互動流程如圖 3-1 所示，主要可分為以下三個階段：

(1) 輸入階段：使用者（如學生、律師）以自然語言提問。系統首先會自動識別詰問階段（如主詰問、反詰問），並運用分類規則判定問題類型。

(2) AI 應答階段：系統根據專屬提示詞（Prompt）與檢索增強生成（RAG）技術，調用知識庫與語用規則。AI 證人會依據問題類型、提問意圖及法庭情境，生成高度擬真的答覆。

(3) 輸出階段：系統即時以帶有語調、停頓變化的語音呈現AI證人的回應，模擬真實法庭中證人緊張、遲疑或堅定的口氣，以確保學習者能沉浸於真實的互動情境。

“

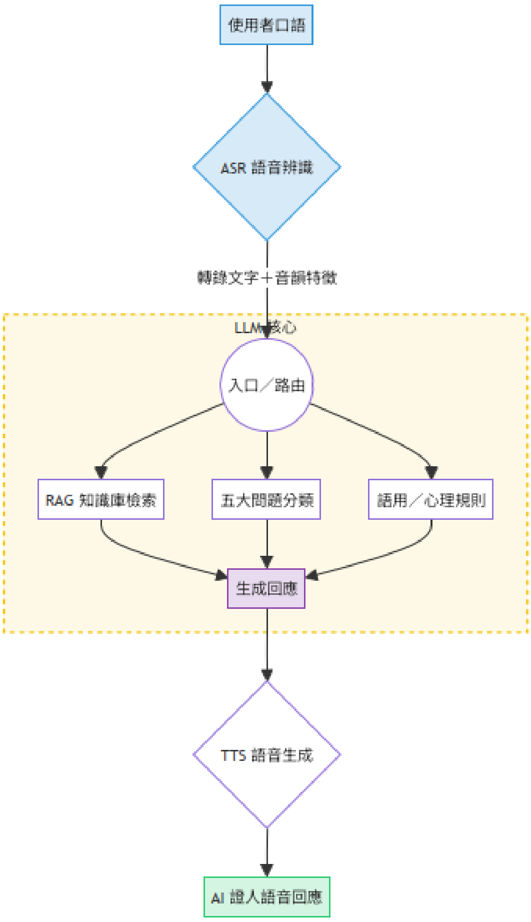


圖 1：法庭詰問 AI 模擬系統之架構與互動流程

本圖呈現使用者從發出口語提問，到接收 AI 證人語音回應的完整處理流程。使用者的語音輸入先經由 ASR 模組轉換為文字，接著進入「LLM 核心」進行處理。此核心整合了 RAG 知識庫檢索、五大問題分類與語用—心理規則三大元件，以生成兼具專業性與行為擬真度的回應文本。最後，文本再經由 TTS 模組轉換為帶有語氣變化的語音，完成一次詰問互動。

3.2 三大核心架構

為讓 AI 模擬證人，實現「行為擬真度」，本系統由三大核心架構構成：

3.2.1 語用-心理判斷規則模組

本模組結合法庭語用學與認知心理學理論 (Cotterill, 2003; Coulthard and Johnson, 2017)，使 AI 能在壓力、誘導或澄清等情境下展現遲疑、拒答或修正等行為，避免傳統對話系統「過度流暢」「過度迎合」而失真的問題。其核心為大規模提示詞工程（Prompt Engineering），涵蓋角色設定、行為規則、語用風格與情境化策略，全文逾萬字：主詰問中設計為誠懇配合，反詰問則轉為謹慎閃避，並在誘導性提問下展現「拒答」或「模糊化」的自然反應。由於提示詞篇幅龐大且細節繁複，若全文刊載將嚴重影響篇幅與可讀性，因此本文僅呈現設計原則與部分範例。相較於僅關注正確性與流暢度的既有方法 (Ji et al., 2023; Belz and Kow, 2011)，本研究更強調在高壓法庭場景中重現臨場語用與行為反應。為此，本模組採取 可控文本生成 (Controllable Text Generation) 的設計路徑(Hu et al., 2017)，將語言學與心理學理論轉化為結構化提示詞的 meta-rules。例如，「反詰問」標籤會觸發簡短句式與不確定詞彙（如「也許」「不一定」），並避免主動補充，從而確保生成結果符合專業語境的「行為擬真度」。

3.2.2 五大問題分類機制

為充分重現真實法庭詰問現場，系統將所有提問自動歸類為以下五大類型，五大問題分類設計，主要參考 Cotterill(2003)對於法庭詰問話語結構與敘事權力的分類分析，同時融入本研究者在台灣模擬法庭教學與檢察官公訴實務經驗，強化在本土法庭詰問語境下的實用性與臨場擬真度。每類問題均訂有專屬語用規則，並可依主詰問/反詰問進一步細緻調整，問題共分為五類，詳細範例放在附錄 A：

(1) 背景問題（Background Questions）

此類問題重在建立證人專業信任，故要求 AI 證人能以冷靜、自信、條理清晰之口吻，簡要說明自身資格與經驗

(2) 證據問題（Evidence Questions）

此類問題是法庭詰問的核心，發問者希望證人基於本案事實或證據回答問題，因此，此階段的應答規則以抑制模型產生學界所稱的「幻覺」(Hallucination）為首要目標(Ji et al., 2023)，嚴格限制其回應不得脫離卷證範圍或憑空捏造。

(3) 推論或專業判斷問題（Inference Questions）

此類問題在專家證人的詰問中非常常見，發問者希望證人做出推論，故常以：「你認為..」、「依照你的經驗…」做為問題開頭。AI 證人對此類問題的回答必須基於專業及本案證據回答。

(4) 澄清問題（Clarification Questions）

此類問題在詰問中亦很常見。在詰問中，發問者會對證人已經回答過的問題繼續追問；或針對關鍵事實希望證人補充；或希望專家證人對專業術語以白話說明或舉例解釋。

(5) 題組問題（Grouped Questions）

此類問題通常包含許多子題，發問者是以題組方式進行詰問，此時，證人不一定要每一題都詳細回答，但必須維持邏輯一致性（Self-Consistency），且必須在後續回答中適時補充前一問題的細節，以建構完整的證詞敘事。

3.2.3 本土卷宗知識庫與 RAG 檢索技術應用

本系統的知識核心來自去識別化與結構化處理後的臺灣真實刑案卷宗，目前涵蓋 2 件案件與 4 個角色（被告、目擊證人、被害人、專家證人）。在技術上，我們利用 OpenAI GPTs (OpenAI, 2023) 的知識庫檢索功能，將卷宗切分為片段並轉換為向量嵌入，以便檢索。使用者提問後，系統會檢索最相關片段並輸入大型語言模型生成回應，屬於典型的檢索增強生成 (Retrieval-Augmented Generation, RAG) 範式 (Lewis et al., 2020)，能降低「幻覺」(Ji et al., 2023) 並確保回答有所依據。

4 實驗設計與成果分析

本章透過專為法律專業優化的「專家圖靈測試」，系統性評估本模擬系統之行為擬真度。該測試借鑒圖靈「模仿遊戲」(Turing, 1950)，但在評審組成、評量指標與評分方式上均針對法律專業場域進行調整。由於行為擬真度是一個多維概念，本研究將其拆解為四個核心面向，以全面檢視 AI 證人的表現：

(1) 專業內容準確度 (Professional Content Accuracy)：以事實一致性為基礎，確保 AI 回答忠於卷宗事實 (Ji et al., 2023)。

(2) 情境適應 (Situational Adaptability)：指 AI 能否根據上下文與詰問策略，展現如真人般的靈活反應 (Deriu et al., 2021)。

(3) 人味化 (Human-likeness)：評估 AI 是否能模擬真人的語言特徵與「恰當的不完美」(Celikyilmaz et al., 2020)。

(4) 邏輯一致性 (Logical Consistency)：要求 AI 在長詰問中，維持證詞前後一致 (Dziri et al., 2022)。

透過對這四個面向的綜合評鑑，本研究希望突破傳統僅重視流暢度的評估框架，推動 AI 評估從「能答」到「像真人在答」的典範轉型。為此，我們邀集具法律專業與一般背景的評審，以盲評方式進行檢驗。

4.1 實驗目的

基於前述四個評估面向，本研究的實驗目的如下：

(1) 檢驗 AI 證人之行為擬真度：評估系統在四大面向上的綜合表現。

(2) 驗證本研究框架的有效性：透過比較「客製化 GPT」與「GPT-Vanilla」基線，驗證本研究核心架構的貢獻。

(3) 評估本研究框架的穩健性：將同一框架應用於不同 LLM (GPT-4.1 與 Gemini 2.5 Flash)，評估其跨模型的潛力。

4.2 實驗設計與資料來源

為確保實驗的嚴謹與可重現性，本前導研究採取下列設計：

(1) 案件選擇與角色設定：本研究選用臺灣新北地方法院「111 年第 3 場國民法官模擬法庭」之「李家銘殺人案」卷宗為知識依據，設定 AI 證人為該案法醫鑑定人（化名「張開平」）。需說明的是，作為前導研究，本次評估僅聚焦於單一案件，其結論的通用性待後續研究擴充。

(2) 實驗材料生成：建立四組筆錄作為比較對象。(a) 真人筆錄：由專業書記官紀錄的模擬法庭問答，作為黃金標準 (Gold Standard, GS) (Lin, 2004)；(b) 基於 GPT-4.1 的客製化系統筆錄（後文簡稱『客製化 GPT』）；(c) 套用相同框架的客製化 Gemini2.5 Flash 筆錄（後文簡稱『客製化 Gemini』）。(b)(c)兩者是完全相同的資料庫與提示詞條件；另加入(d) GPT-Vanilla（同一 GPT-4.1、但提示詞中僅僅提供基礎的角色設定（『你是一位名叫張開平的

法醫』）與 RAG 檢索的卷宗內容，而未加入任何關於語氣、風格、應對策略的引導。）作為香草基準。三組 AI 筆錄都是回答與真人筆錄相同的 50 題詰問（涵蓋主詰問與反詰問）。雖本系統具備完整語音功能，但為確保評估的標準化與可重現性，本次評估全程採用書面文字筆錄進行。

(3) 評審組成與盲評程序：邀集五位評審（n=5），其中三位具法庭經驗的法律專家，兩位為一般大學畢業者。四組筆錄以隨機順序呈現，並移除所有可能透露來源的線索（如模型名稱、格式化語言），避免風格偏誤。評審需先辨識出真人筆錄，並將其設定為本次實驗的 GS（各指標自動計為 5.0 分），以確立基準分數。此設計並非旨在衡量 AI 與真人間的差異，而是確立一個明確且一致的基準分數，作為 AI 系統擬真度比較的參照點。

(4) 評分方式：針對三組 AI 筆錄，評審使用五點 Likert 量表（1 = 完全不像真人，5 = 幾可亂真），並依四個面向——專業內容準確度、情境適應 / 臨場應變、人味化及邏輯一致性——進行評分，同時並提供質性回饋。

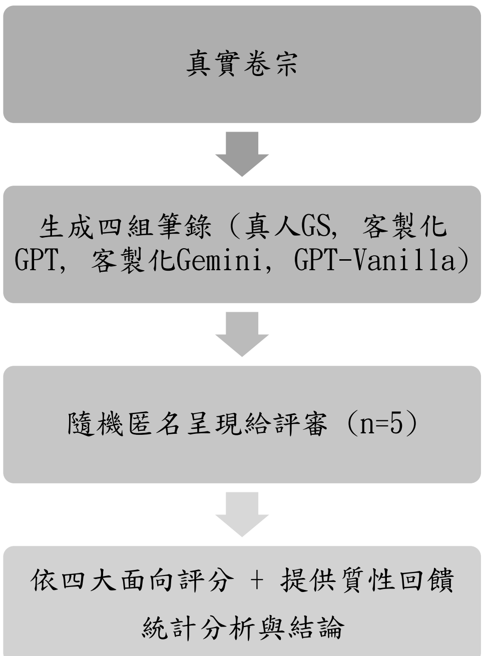


圖 2：專家圖靈測試之評估流程圖

4.3 行為擬真度評量與成果分析

本節呈現專家圖靈測試的量化數據與質性回饋分析。為確保評估的客觀性，我們進行了以下統計檢驗：

(1) 跨評審一致性：採用 Kendall’s W 檢定來衡量五位評審之間的一致性。

(2) 模型間比較：採用 Wilcoxon 符號等級檢定（n=5, α=.05）來比較「客製化 GPT」與「GPT-Vanilla」基線之間的差異。

整體擬真度總評分與結論

根據評審的綜合評分，採用本研究行為規訓框架的 AI 系統，其擬真度顯著優於未經規訓的基線。其中，「客製化 GPT」的總平均分數為 4.55 分（見表 2），表現尤其出色。「客製化 Gemini」的表現同樣穩健，總平均分為 4.30 分。

詳細的分項評分結果如表 2 所示。從表中可見，本研究的框架在「情境適應」與「人味化」兩個維度上帶來了最顯著的提升。「客製化 GPT」在這兩項的得分（均為 4.6 分）顯著高於「GPT-Vanilla」基線（3.0 及 2.9 分），此差異達到了統計上的顯著性（$p < .05$），與表格中的星號（*）標記一致。

行為細節分項評分與定性回饋

為進一步解析差異成因，評審對各項細節指標進行了 1-5 分的李克特量表評分。結果如下表 2 所示：

表 2：專家圖靈測試評分表。

(評分由五位評審進行，採五點李克特量表)

評量面向	真人筆錄 (M±SD)	客製化 GPT (M±SD)	客製化 Gemini (M±SD)	GPT-Vanilla (M±SD)
1.內容準確度	5.0± 0.0	4.4± 0.3	4.4± 0.4	4.1± 0.4
2.情境適應	5.0± 0.0	4.6± 0.3 *	4.2± 0.5	3.0± 0.5
3.人味化	5.0± 0.0	4.6± 0.2 *	4.0± 0.6	2.9± 0.2
4.邏輯一致	5± 0.0	4.6± 0.3	4.6± 0.2	4.1± 0.3
總平均	5.0	4.55	4.30	3.53

註：統計分析以 Python 3.10／R 4.3 進行，可重現。數值為 5 位評審（3 位法律專家，2 位一般參與者）Likert 評分之平均 ± 標準差；總

平均為四面向等權平均（不另計 SD）。跨評審一致性以 Kendall’s W 驗證。* 表示客製化 GPT 系統與 GPT-Vanilla 基線在該維度上的差異，經 Wilcoxon 符號等級檢定達到統計顯著性 ($p < .05$)。本研究的重點比較在於驗證客製化系統相較於香草基線的改進幅度，詳細的效應量與統計值於 4.4 節中闡述。

評審定性回饋與分析

在質性回饋中，評審普遍認為兩個模型在專業內容準確度與邏輯一致性上均表現出色，GPT-Vanilla 表現一般。然而，細微差異體現在：客製化 Gemini 在提供「2.5 公分」等具體數據上更接近真人筆錄的細緻度；而本研究的客製化 GPT 則在情境適應與人味化上顯著勝出，這也與其和香草基準之間最大的分數差距（+1.60 及 +1.70）相吻合。評審特別指出，客製化 GPT 能以「不一定」、「這其實跟角度有關」等更靈活、口語化的方式應對誘導性詰問，並透過大量語助詞與自然的停頓，成功模擬了真人在壓力下的語言行為，甚至在回答中出現了輕微的細節重複或回溯，這些「不完美」反而讓它聽起來像一個活生生、有血有肉的人在發言。更貼近本研究「行為擬真度」的核心目標。不過，有評審補充，客製化 Gemini 在面對帶有錯誤資訊的誘導性提問時，能更直接且準確地予以糾正，展現了另一種形式的應變能力。

4.4 綜合討論與研究限制

綜合量化數據與質性回饋，本研究的成果初步表明，經由專業知識與語用規則客製化的 AI 證人，已能在行為擬真度上展現潛力。如前節所述，本研究框架在「情境適應」與「人味化」兩個維度上帶來了最顯著的提升。

評審的質性回饋為此提供了更深入的解釋。有評審特別指出，客製化 GPT 能以「不一定」、「這其實跟角度有關」等更靈活、口語化的方式應對誘導性詰問，並透過大量語助詞與自然的停頓，成功模擬了真人在壓力下的語言行為。這些「不完美」的細節反而讓它聽起來更像一個真實的人，這與本研究追求「行為擬真度」的核心目標相符 。

然而，本研究亦有其限制：

(1) 案例範圍：實驗僅基於單一刑案卷宗，其結論在跨案件類型上的通用性仍待後續驗證 。

(2) 評估媒介：雖語音互動是本系統的核心創新，但本次為求標準化與可重現性，評估乃基於書面筆錄進行，未能涵蓋語氣、語速等重要的非語言線索 。

(3) 樣本規模：本研究的評審團人數（n=5）有限，因此其量化結果應被視為探索性的初步發現，有待未來更大規模的評估來證實 。

5. 結論與未來展望

5.1 研究結論

本研究的實驗結果表明，所提出的「行為擬真度」框架在法律教育輔助工具的開發上具有潛力。作為一項前導研究 (pilot study)，其成果為後續更大規模的研究奠定了基礎。與過往僅依賴文字互動的模擬系統相比，本研究首次將語音模擬納入本土法庭詰問練習，讓使用者得以進行「口語提問—語音回應」的雙向演練，此特點大幅提升了訓練的沉浸感與真實感。

本研究的主要貢獻可歸納如下：

(1) 提出並操作化「行為擬真度」作為評估核心：補足了既有 AI 評估偏重內容正確性的不足，將臨場語用與行為一致性納入評估框架。

(2) 建構一可重現的跨域系統架構：整合了語用—心理規則、五大問題分類機制與本土卷宗知識庫，以支援高擬真度的口語互動演練。

(3) 以專家圖靈測試框架完成前導實證：透過比較客製化 GPT 與 GPT-Vanilla 基線，驗證了本研究框架對提升擬真度的有效性；並提供了與另一主流模型的敘述性對照，初步展現了方法的潛力。

誠然，如 4.4 節所詳述，本研究作為前導研究，在案例範圍、評估媒介與評審規模上仍有其限制。然而，研究成果仍清晰地驗證了「行為擬真度」作為評估指標的可行性，並展現了生成式 AI 作為可規模化法律培訓工具的巨大潛力。

5.2 未來展望

奠基於本次研究的成果與前述限制，未來的研究可朝以下方向深化：

(1) 深化行為擬真度 (Deepening Behavioral Fidelity)：針對 AI 在壓力下的微行為（如自然遲疑、語助詞）進行更深入的建模，以克服「數位恐怖谷」效應。

(2) 整合多模態評估 (Multimodal Evaluation)：在未來研究中，將評估從文字筆錄擴展至包含語氣、語速等完整多模態互動，以更全面地檢驗擬真度。

(3) 擴充案例與驗證通用性 (Expanding Case Base and Validating Generalizability)：持續擴大本土案例的類型與數量，並招募更多元的評審樣本，以提升系統的通用性與研究結論的外部效度。

(4) 開發自動化評估指標 (Developing Automated Metrics)：探索半自動化指標（如以 NLI 模型檢測邏輯一致性）的可行性，以輔助專家評分，提升評估效率與規模。

參考文獻(References)

Barrows, Howard S. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Academic Medicine, 68(6):443–451. DOI: 10.1097/00001888-199306000-00002

Belz, Anja and Ehud Kow. 2011. Discrete vs. continuous rating scales for language evaluation: Revisiting the human evaluation of NLG systems. In Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), pages 98–102. Association for Computational Linguistics. ACL Anthology ID: P11-2040

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. arXiv preprint arXiv:2006.14799. https://arxiv.org/abs/2006.14799

Chen, Zihan, Wei Wang, Fan-Keng Sun, Hong-Han Shuai, and Wen-Chih Peng. 2025. AgentCourt: A multi-agent simulation framework for court debates. In Proceedings of the 39th AAAI Conference on Artificial Intelligence.

Cotterill, Janet. 2003. Language and Power in Court: A Linguistic Analysis of the O.J. Simpson Trial. Palgrave Macmillan.

Coulthard, Malcolm and Alison Johnson. 2017. An Introduction to Forensic Linguistics: Language in Evidence (2nd ed.). Routledge.

Deriu, Jan, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Gorka Azkune. 2021. A survey on the evaluation of dialog systems. arXiv preprint arXiv:2106.01254. https://arxiv.org/abs/2106.01254

Dziri, Nouha, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating the factual consistency of large language models through natural language inference. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9716–9741. Association for Computational Linguistics. DOI: 10.18653/v1/2022.emnlp-main.663 (ACL Anthology ID: 2022.emnlp-main.663)

Gudjonsson, Gisli H. 2003. The Psychology of Interrogations and Confessions: A Handbook. Wiley.

Hendrycks, Dan, Collin Burns, Steven Basart, et al. 2021. Measuring massive multitask language understanding. In Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2009.03300

Hu, Zhiting, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In Proceedings of the 34th International Conference on Machine Learning (ICML), pages 1587–1596. DOI: 10.48550/arXiv.1703.00955

Ji, Zhijing, Nayun Lee, et al. 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38. DOI: 10.1145/3571730

Lewis, Patrick, Ethan Perez, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pages 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Liang, Percy, Rishi Bommasani, et al. 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110. https://arxiv.org/abs/2211.09110

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In Proceedings of the ACL-04 Workshop on Text Summarization Branches Out, pages 74–81. Association for Computational Linguistics. ACL Anthology ID: W04-1013

Mauet, Thomas A. 2017. Trial Techniques and Trials (10th ed.). Wolters Kluwer.

Mori, Masahiro. 2012. The uncanny valley. IEEE Robotics & Automation Magazine, 19(2):98–100. DOI: 10.1109/MRA.2012.2192811

OpenAI. 2023. GPTs. https://platform.openai.com/docs/gpts. Accessed: October 16, 2025.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02), pages 311–318. DOI: 10.3115/1073083.1073135

Serra, Kevin D. 2024. MootMentorAI: A generative AI-powered platform for law school moot court training. arXiv preprint arXiv:2403.09033. https://arxiv.org/abs/2403.09033

Slater, Mel and Maria V. Sanchez-Vives. 2016. Enhancing our lives with immersive virtual reality. Frontiers in Robotics and AI, 3:74. DOI: 10.3389/frobt.2016.00074

Stanford CodeX and Three Crowns LLP. 2025. CodeX and Three Crowns LLP announce AI cross-examination tool. https://law.stanford.edu/codex-the-stanford-center-for-legal-informatics/codex-and-three-crowns-llp-announce-ai-cross-examination-tool/. Accessed: October 16, 2025.

Turing, Alan. 1950. Computing machinery and intelligence. Mind, 59(236):433–460. DOI: 10.1093/mind/LIX.236.433

附錄 A：五類問題範例

(1) 背景問題（Background Questions）

例：

問：請問你擔任法醫多久？

AI 答：我在法醫所任職 20 年，主要負責

(2) 證據問題（Evidence Questions）

例：

問：請解釋死者死因

AI 答：請看解剖報告第 X 頁，死者死因是

(3) 推論或專業判斷問題（Inference Questions）

例：

問：一般人肋骨間距大概是多少？

AI 答：一般來說，肋骨的間距大約是 2 到 3 公分，但還是因人而異。

(4) 澄清問題（Clarification Questions）

例：

問：請說明什麼是單一銳器穿刺傷？

AI 答：就是由刀子等具穿刺功能的銳器刺入人體所造成的創

(5) 題組問題（Grouped Questions）

例：

問 1： 死者死因是什麼？

AI 答：是單一銳器穿刺傷造成的。

問 2：什麼是單一銳器穿刺傷？

AI 答：就是由刀子等具穿刺功能的銳器刺入人體所造成的創傷，我記得在解剖大體的時候，我們發現死者的右側胸部有一個單一的銳器傷。這個傷口是從右下往左上的方向 ，直接刺穿了橫膈膜。這就是造成心包膜囊裡面大量出血的原因。

Multimodal Approaches for Stress Recognition: A Comparative Study Using the StressID Dataset

Chia-Yun Lee
The Department of Data Science at Soochow University, Taiwan
jessicaleej0033@gmail.com

Matúš Pleva
Department of Electronics and Multimedia Communications, Technical University of Košice, Slovakia
matus.pleva@gmail.com

Daniel Hládek
Department of Electronics and Multimedia Communications, Technical University of Košice, Slovakia
daniel.hladek@tuke.sk

Ming-Hsiang Su
The Department of Data Science at Soochow University, Taiwan
huntfox.su@gmail.com

Abstract

Mental health concerns have garnered increasing attention, highlighting the importance of timely and accurate identification of individual stress states as a critical research domain. This study employs the multimodal StressID dataset to evaluate the contributions of three modalities—physiological signals, video, and audio—in stress recognition tasks. A set of machine learning models, including Random Forests (RF), Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), and K-Nearest Neighbors (KNN), were trained and tested with optimized parameters for each modality. In addition, the effectiveness of different multimodal fusion strategies was systematically examined. The unimodal experiments revealed that the physiological modality achieved the highest performance in the binary stress classification task (F1-score = 0.751), whereas the audio modality outperformed the others in the three-class classification task (F1-score = 0.625). In the multimodal setting, feature-level fusion yielded stable improvements in the binary classification task, while decision-level fusion achieved superior performance in the three-class classification task (F1-score = 0.65). These findings demonstrate that multimodal integration can substantially enhance the accuracy of stress recognition. Future research directions include incorporating temporal modeling and addressing data imbalance to further improve the robustness and applicability of stress recognition systems.

Keywords: Stress Detection, Multimodal Machine Learning, Audio-Visual Features, Feature-Level Fusion.

1 Introduction

As the pace of modern society accelerates and life pressures intensify, mental health is getting more attention. The World Health Organization (WHO) designates October 10th each year as World Mental Health Day, emphasizing that mental health is a fundamental human right and urging all sectors to address psychological issues and provide necessary resources. However, in high-pressure environments, many individuals struggle to recognize and manage stress, which can gradually accumulate and lead to more serious mental health challenges.

Stress is essentially a state, both mental and physical, that happens when people feel the demands of their environment are beyond their ability to cope, threatening their well-being (Lazarus, R.S. et al., 1984). It is a dynamic and interactive process that involves the individual's cognitive appraisal and coping strategies in response to stressors. Research has shown that stress has both direct and indirect effects on mental health, particularly through the regulation of psychological states via negative emotions (Moreta-Herrera, R., et al, 2023). While moderate stress can foster adaptation and motivation, prolonged and unmanaged stressors may negatively impact the nervous system, mental health, and behavior patterns (Hsu, Y. F., 2021).

Taking the campus as an example, students face multiple pressures from academic work, interpersonal relationships, and future

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 29–34
November 20-22, 2025

development, which often brings their mental health issues into the news spotlight. According to statistical data released by the Ministry of Education's Campus Safety and Disaster Prevention Center in 2024（教育部校安通報中心, 2024）, suicide and self-harm incidents have ranked first in campus safety-related accidental reports for the past three years, accounting for 33% of all reported accidents. The number of deaths in higher education institutions remains high. In recent years, universities have begun implementing mental health leave, believing that it helps students with self-awareness and provides an opportunity for short-term adjustment, hoping to reduce the incidence of such incidents.

Currently, the assessment of psychological stress primarily relies on traditional questionnaire-based surveys (Scale, P.S., 1983). However, these methods are limited by their high subjectivity and lack of real-time responsiveness, which hinder the implementation of timely intervention strategies. Therefore, developing an objective and real-time stress monitoring technology has become a crucial research direction. Furthermore, existing research and datasets on stress detection face notable limitations, including small dataset sizes, a lack of diverse stress sources, and unimodal data constraints. To address these issues and advance the field of stress recognition, this study will utilize the rich resources of the StressID dataset (Chaptoukaev, H., et al., 2023). It aims to optimize and evaluate the performance of various unimodal and multimodal fusion models, with the goal of developing more objective and reliable stress identification techniques that can enhance mental health monitoring and intervention capabilities.

2 Related Literature

2.1 Stress Recognition Research

With the growing awareness of mental health, recent years have witnessed increasing research efforts dedicated to enhancing the accuracy of stress detection through a wide range of features and classification models. One notable contribution is the WESAD dataset introduced by Schmidt et al. (2018), which integrates multiple wearable sensor signals with emotion annotations and has since become a widely used benchmark for developing and evaluating multimodal stress recognition systems. Building on this resource, Abdelfattah et al. (2025) conducted a comparative analysis of machine learning and deep learning models using the WESAD dataset. Their findings suggest that deep learning methods provide superior cross-subject generalization but are computationally demanding, limiting their feasibility for real-time applications. In contrast, traditional machine learning models demonstrate greater computational efficiency and achieve high accuracy in personalized settings—reaching up to 99.8% F1 score—yet they suffer from limited generalizability. To address these shortcomings, ensemble learning has been highlighted as a promising strategy for enhancing both robustness and generalization in stress recognition. Extending this line of research, the present study explores the StressID multimodal dataset, with particular emphasis on evaluating the contributions of different modalities and investigating the impact of fusion strategies on model performance.

2.2 Classification Models for Stress Detection

To achieve this, a set of established machine learning and deep learning models is considered. Random Forest (RF) is an ensemble learning approach composed of multiple decision trees that improves classification stability and accuracy by employing a voting mechanism for both classification and regression tasks. Its performance depends on hyperparameters such as the number of estimators (n_estimators), the splitting criterion (criterion), and the maximum tree depth (max_depth), which are generally optimized through cross-validation. Support Vector Machine (SVM) is a supervised learning model that identifies the optimal hyperplane separating data points of different classes with the maximum margin, making it effective for classification tasks with well-defined decision boundaries. Its effectiveness relies on the selection of the kernel function, the regularization parameter (C), and the kernel coefficient (gamma). K-Nearest Neighbors (KNN) is a non-parametric, distance-based algorithm that classifies new data points by identifying the K nearest neighbors and applying majority voting, with hyperparameters including the number of neighbors (n_neighbors), the weighting scheme, and the neighbor computation algorithm. Multilayer Perceptron (MLP), a feedforward artificial neural network, is capable of modeling complex nonlinear relationships through an input layer, one or more hidden layers, and an output layer. Its performance is shaped by factors such as the activation function, learning rate, optimization algorithm, and hidden layer configuration. Finally, the Deep Belief Network (DBN), composed of stacked Restricted Boltzmann Machines (RBMs), is a deep generative model that performs unsupervised pretraining to capture hierarchical data representations, followed by supervised fine-tuning for classification. DBNs are particularly valued for their strong feature extraction capabilities, especially in handling structured and high-dimensional data.

3 Dataset Collection and Processing

3.1 Dataset Description

This study employs the StressID dataset, a multimodal resource integrating physiological signals, video, and audio recordings. Figure 1 illustrates the structure of the dataset. Data collection followed a rigorous and reproducible experimental protocol comprising 11 tasks organized into four main blocks: guided breathing, emotional video clips, seven interactive stress-inducing tasks, and a relaxation phase. These diverse tasks were designed to elicit varying stress responses among participants.

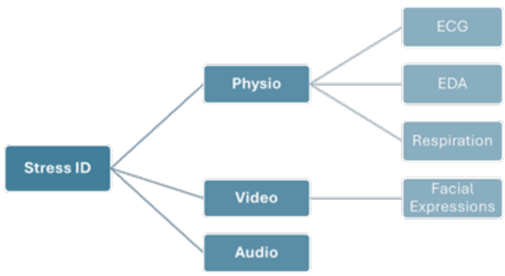


Figure 1: Multimodal Structure of the StressID Dataset

The experiment involved 65 healthy adult participants. Following each task, participants completed self-assessment questionnaires to report their perceived stress, relaxation, valence, and arousal levels. In this context, valence refers to the positive or negative emotional state experienced during a task, whereas arousal reflects emotional activation or engagement. These self-reported measures were subsequently used for supervised learning models to generate binary labels (stressed vs. not stressed) and ternary labels (relaxed, neutral, stressed).

All multimodal data were collected synchronously and processed through task segmentation and annotation procedures. The final StressID dataset comprises over 39 hours of annotated recordings, including 711 physiological signal recordings, 587 video segments, and 385 audio recordings. Its scale and diversity make the dataset one of the most extensive publicly available stress identification resources suitable for unimodal and multimodal research.

3.2 Dataset Processing

The StressID dataset provides baseline stress classification models in unimodal and multimodal settings. This section describes the feature extraction and preprocessing procedures for the three unimodal data types, inputs for subsequent machine learning models. For physiological signals, 35 features were extracted from the electrocardiogram (ECG), 23 from electrodermal activity (EDA), and 40 from respiration signals. All signals were first processed using a Butterworth filter to reduce high-frequency noise and baseline drift. Extracted features include statistical and physiological measures such as heart rate variability (HRV), skin conductance level (SCL), skin conductance response (SCR), and respiratory rate variability (RRV), all intended to quantify the participants' physiological states.

Video data were processed using the OpenFace library to extract facial features, including Action Units (AUs) and eye gaze trajectories. These features' mean and standard deviation were calculated to capture facial expressions and gaze dynamics, resulting in an 84-dimensional feature vector for each video segment. Audio recordings were down-sampled to 16 kHz, and amplitude-based Voice Activity Detection (VAD) was applied to remove non-speech segments. Handcrafted features were extracted, including Mel-frequency cepstral coefficients (MFCCs) and their derivatives, spectral centroid, and other spectral features, forming a 114-dimensional feature vector. Additionally, speech embeddings were obtained from the pre-trained Wav2Vec 2.0 (W2V) model. Embeddings were extracted every 20 milliseconds and averaged over time to generate a 513-dimensional representation per utterance, capturing variations in frequency, energy, and speech rhythm.

All features, except those extracted by Wav2Vec 2.0 (which were classified using a linear classifier), were standardized and used as inputs to machine learning models, including Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and K-Nearest Neighbors (KNN). These models were trained and evaluated under various parameter configurations to predict stress-related labels, including binary and three-class classification.

Multimodal integration strategies were also explored to improve classification performance. The first approach, feature-level fusion, concatenates features from each modality into a single high-dimensional vector, which is then used as input to machine learning models. The second approach, decision-level fusion, trains independent models for each modality and combines their predictions using ensemble rules such as summation, product, averaging, or maximum to generate the final decision.

A notable challenge in the StressID dataset is class imbalance, particularly in audio data, as speech tasks are often associated with elevated stress levels. SMOTE was applied to balance binary-class audio data and the multimodal training set to address this. However, in three-class audio classification, the "relaxation" category is underrepresented due to the limited presence of audio during relaxation tasks. The scarcity of relaxed audio samples and the absence of characteristic relaxed speech features limit the effectiveness of resampling in this scenario.

3.3 Model Performance Evaluation

To assess the classification performance of the model in stress detection tasks, this study evaluated the model on the test dataset using weighted F1-score and balanced accuracy. The evaluation metrics are defined as follows:

$$F1_{weighted} = \sum_{i=1}^{n} w_i \times F1_i \quad (1)$$

$$Balanced\ Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i} \quad (2)$$

These metrics were used to measure the model's performance in both binary and multi-class stress classification tasks across different modalities. The weighted F1-score emphasizes classification accuracy while taking the class distribution into account. On the other hand, balanced accuracy mitigates the influence of class imbalance by averaging the recall across all classes, providing a fairer assessment of the model's ability to recognize each class equally.

4 Experimental Results and Discussion

This study implements various classification models using Python's scikit-learn library, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Multi-layer Perceptron (MLP), with parameter tuning for comparison. The MLP model utilizes multiple combinations of hidden layers, SVM employs the RBF kernel with C-value adjustment and a fixed gamma of 0.00714, KNN investigates the effect of different numbers of neighbors, and RF investigates the effect of different tree depths. All models were evaluated using 10 random splits to ensure robustness and reliable performance estimation.

As shown in Table 1 and Table 2, for the binary-stress classification task, the Physio modality with Random Forest achieved the best performance (F1 = 0.751) with a maximum tree depth of 20. For the three-class affect classification task, the Audio modality with SVM performed best (F1 = 0.577) with a C-value of 10. Overall, all models performed better on the binary classification task, with the Physio modality demonstrating the best binary performance, while the Audio modality outperformed others in the three-class setting.

Table 1. Comparison of unimodal baseline performances on the binary-stress classification task.

	Binary-stress	
Classifier	**F1-score**	**Accuracy**
Video. AUs + RF	0.702±0.03	0.703±0.03
Video. AUs + SVM	0.701±0.03	0.701±0.02
Video. AUs + KNN	0.706±0.03	0.705±0.03
Video. AUs + MLP	**0.708±0.04**	**0.708±0.04**
Audio. HC features + RF	0.689±0.07	0.629±0.07
Audio. HC features + SVM	0.713±0.05	0.664±0.05
Audio. HC features + KNN	0.576±0.04	0.627±0.03
Audio. HC features + MLP	**0.718±0.07**	**0.671±0.07**
W2V 2.0 classifier + MLP	**0.725±0.05**	**0.667±0.05**
Physio. HC features + RF	**0.751±0.03**	**0.744±0.03**
Physio. HC features + SVM	0.733±0.03	0.725±0.03
Physio. HC features + KNN	0.696±0.04	0.689±0.04
Physio. HC features + MLP	0.712±0.03	0.709±0.03

Table 2. Comparison of unimodal baseline performances on the affect3-class classification task.

Classifier	Affect3-class F1-score	Affect3-class Accuracy
Video. AUs + RF	0.557±0.05	0.555±0.05
Video. AUs + SVM	**0.565±0.03**	**0.559±0.03**
Video. AUs + KNN	0.563±0.04	0.558±0.04
Video. AUs + MLP	0.564±0.03	0.557±0.04
Audio. HC features + RF	0.515±0.07	0.478±0.06
Audio. HC features + SVM	**0.577±0.04**	**0.535±0.06**
Audio. HC features + KNN	0.526±0.06	0.491±0.07
Audio. HC features + MLP	0.558±0.03	0.519±0.07
W2V 2.0 classifier	**0.625±0.05**	**0.564±0.05**
Physio. HC features + RF	0.569±0.02	0.565±0.02
Physio. HC features + SVM	**0.576±0.04**	**0.574±0.04**
Physio. HC features +KNN	0.561±0.02	0.552±0.03
Physio. HC features + MLP	0.537±0.04	0.53±0.04

In the multimodal analysis, three approaches are covered: unimodal models, feature-level fusion, and decision-level fusion. Additionally, various classifiers (SVM, RF, MLP, KNN) are compared.

According to Table 3, the best unimodal performance is achieved by Audio + SVM (F1 = 0.73), with parameters C = 10. Among the fusion strategies, feature-level fusion with MLP (1 hidden layer, 100 units) or SVM (C = 1.0, gamma = 0.00714) achieved the best performance in the binary-stress task (F1 = 0.72), slightly outperforming the decision-level fusion results.

Table 3. Comparison of multimodal baseline performances on the binary-stress classification task.

Classifier	Binary-stress F1-score	Binary-stress Accuracy
Video. + SVM	0.7±0.04	0.64±0.05
Audio. + SVM	**0.73±0.02**	**0.68±0.02**
Physio. + RF	0.71±0.04	0.63±0.04
Feature level fusion + MLP	**0.72±0.06**	**0.66±0.07**
Feature level fusion + KNN	0.61±0.07	0.63±0.07
Feature level fusion + RF	0.67±0.05	0.57±0.03
Feature level fusion + DBN	0.63±0.05	0.57±0.04
Feature level fusion + SVM	**0.72±0.06**	**0.66±0.06**
RF + Sum level fusion	0.72±0.03	0.65±0.03
RF + Product level fusion	0.72±0.03	0.64±0.03
RF + Average level fusion	0.72±0.03	0.65±0.03
RF + Maximum level fusion	0.72±0.04	0.63±0.04

In contrast, in the affect3-class task in Table 4, the multimodal fusion strategies clearly outperform the unimodal models. Among them, the Decision-level fusion with RF (max_depth = 25, random_state = 0) + Sum/Average achieved the best performance, with F1 = 0.65. Feature-level fusion with MLP (F1 = 0.62) also showed a close performance, demonstrating practical potential.

Table 4. Comparison of multimodal baseline performances on the affect3-class classification task.

Classifier	Affect3-class F1-score	Affect3-class Accuracy
Video. + SVM	**0.58±0.06**	**0.55±0.06**
Audio. + SVM	0.52±0.06	0.49±0.04
Physio. + RF	0.52±0.05	0.5±0.06
Feature level fusion + MLP	**0.62±0.05**	**0.61±0.04**
Feature level fusion + KNN	0.53±0.04	0.56±0.06
Feature level fusion + RF	0.54±0.06	0.49±0.06
Feature level fusion + DBN	0.34±0.11	0.35±0.04
Feature level fusion + SVM	0.57±0.05	0.51±0.04
RF + Sum level fusion	**0.65±0.06**	**0.6±0.06**
RF + Product level fusion	0.64±0.06	0.6±0.06
RF + Average level fusion	0.65±0.06	0.6±0.06
RF + Maximum level fusion	0.63±0.04	0.59±0.04

Overall, multimodal fusion strategies outperform unimodal models in both tasks. Feature-level fusion is more suitable for the binary-stress task, while Decision-level fusion shows its advantage in the affect3-class task. In comparison, KNN and DBN underperformed overall, with both accuracy and stability being relatively low.

5 Conclusion and Future Work

This study investigated unimodal and multimodal approaches for stress recognition using the StressID dataset. The results demonstrate the effectiveness of multimodal fusion, with feature-level fusion providing stable performance in binary stress classification, while decision-level fusion achieves superior performance in the three-class affective classification task. Despite these promising outcomes, challenges remain, particularly regarding class imbalance. The underrepresentation of the “relaxation” category adversely affects the performance of three-class classification models. Future research should explore strategies to mitigate these imbalances and consider the incorporation of temporal models, such as LSTM, GRU, or Transformer architectures, to better capture the dynamic nature of stress responses. Additionally, further investigation into the feasibility of these models for real-time monitoring and practical deployment is essential to enhance the timeliness, robustness, and overall accuracy of mental health interventions.

References

Lazarus, Richard S., and Susan Folkman. 1984. *Stress, appraisal and coping.* New York: Springer.

Moreta-Herrera, Ricardo, D. Zumba-Tello, J. de Frutos-Lucas, S. Llerena-Freire, A. Salinas-Palma, and A. Trucharte-Martínez. 2023. The role of negative affects as mediators in the relationship between stress and mental health in Ecuadorian adolescents. *Health Psychology Report*, 11(3), 241-251.

Hsu, Yu-Fang. 2021, June. A study of the relationships between perceived stress and the quality of life for clinical nurse preceptors (Master's thesis). Nanhua University, Chiayi County, Taiwan.

教育部校園安全暨災害防救通報處理中心資訊網. 2024, January.「教育部 111 年各級學校校園安全及災害事件分析報告」.

Scale, P. S. (1983). Perceived Stress Scale.

Chaptoukaev, H., V. Strizhkova, M. Panariello, B. D'alpaos, A. Reka, V. Manera, S. Thümmler, E. Ismailova, N. Evans, F. F. Bremond, M. Todisco, M. Zuluaga, and L. M. Ferrari. 2023. StressID: A multimodal dataset for stress identification. In *NeurIPS 2023 - 37th Conference on Neural Information Processing Systems*, NIST, New Orleans, United States.

Philip Schmidt, Attila Reiss, Robert Dürichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. *In Proceedings of the 2018 International Conference on Multimodal Interaction (ICMI '18)*, pages 400–408. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3242969.3242985

Eman Abdelfattah, Shreehar Joshi, and Shreekar Tiwari. 2025. Machine and deep learning models for stress detection using multimodal physiological data. *IEEE Access,* 13:2154–2166. https://doi.org/10.1109/ACCESS.2024.3525459

StressID: A Multimodal Dataset for Stress Identification, Access date: 2025/02/07. https://github.com/robustml-eurecom/stressID?tab=readme-ov-file

Beyond Binary: Enhancing Misinformation Detection with Nuance-Controlled Event Context

Elijah Frederick Albertson, Retnani Latifah, Yi-Shin Chen
National Tsing Hua University, Hsinchu, Taiwan
yishin@gmail.com

Abstract

Misinformation rarely presents itself as entirely true or entirely false. Instead, it often embeds partial truths within misleading contexts, creating narratives that blur the boundary between fact and falsehood. Traditional binary fact-checking frameworks fail to capture this nuance, forcing complex claims into oversimplified categories. To address this gap, we introduce **MEGA**, a multidimensional graph framework designed to classify ambiguous claims, with a particular focus on those labelled "*Somewhat True*." MEGA integrates event evidence, spatio-temporal metadata, and a quantifiable nuance score. Its Event Candidate Extraction (ECE) module identifies supporting or contradicting evidence, while the Nuance Control Module (NCM) injects or removes nuance to assess its effect on classification. Experiments show that nuance is both detectable and learnable: adding nuance improves borderline discrimination, while stripping it leads the decisions toward false extremes and conceals partial truth. Our top model —nuance-injected without score weighting —improve accuracy and F1 score by 15 and 16 points over the claims-only baseline, and 6 and 9 points over the ECE-only variant. These results show that explicitly modeling nuance alongside context is crucial for classifying mixed-truth claims and advancing fact-checking beyond binary judgments.

Keywords: Misinformation detection, Linguistic nuance, Event-guided evidence

1 Introduction

The rapid growth of online media has fueled an overwhelming spread of misinformation (Sharma et al., 2019; Hu et al., 2025a). Because misleading narratives often interweave genuine facts with distortions, separating truth from fiction has become increasingly difficult. Traditional fact-checking pipelines, built on binary true/false labels (Wang et al., 2020a), are ill-suited for claims that fall into the borderline category—especially those tagged *Somewhat True*. Such claims typically contain accurate information that is exaggerated, stripped of context, or paired with omissions (Rashkin et al., 2017), making their classification inherently challenging.

This challenge connects to the notion of certainty, long studied in pragmatics and discourse through phenomena such as epistemic modality, evidentiality, doubt, and hedging (Rubin, 2007). These signals express how confidence is conveyed, and in computational terms can be characterised by polarity (support vs. contradiction) and intensity (strength of stance). Yet, recent work on causal epistemic consistency demonstrates that current language models struggle to remain stable when distinguishing such fine-grained cues (Cui et al., 2025). Motivated by these limitations, we manually analysed 150 *Somewhat True* claims and observed recurring linguistic patterns: hedging markers ("may," "could"), context-sensitive phrasing, and contrastive framing. These are not new facts, but structural signals—indicating that *Somewhat True* is not merely a midpoint between False and True, but a distinct category shaped by nuance.

Building on this observation, we design two key modules. A **Nuance Control Module (NCM)** manipulates hedging and ambiguity markers to probe how linguistic framing influences classification. An **Event Candidate Extraction (ECE)** module retrieves and summarises event-level snippets as exter-

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 35–44
November 20-22, 2025

nal evidence, grounding claims in verifiable context. Together, these modules allow us to test whether nuanced linguistic cues help or hinder borderline judgments, and motivate our inclusion of score-aware evidence that weights semantic, temporal, spatial, and nuance features.

To integrate these signals, we propose the **Multidimensional Event-Guided Analysis Graph (MEGA)**, a graph-based framework that links claims to event evidence and metadata while encoding semantic, temporal, spatial, and nuanced relations. Experimental results show that injecting nuance improves performance in borderline cases: our best configuration, a nuance-injected model without score weighting, achieves a 15-point and 16-point improvement on accuracy and F1 scores over the claims-only baseline. Conversely, removing nuance pushes decisions toward extremes and obscures partial truths. These findings demonstrate that explicitly modelling nuance, alongside contextual evidence, is essential for reliable classification of mixed-truth claims.

The key contributions are:

- **Nuance Control Module (NCM)** — injects or removes hedging, conditional, and ambiguity markers to test framing effects.
- **MEGA** —a configurable graph that links claims to event evidence, metadata, and linguistic nuance features via semantic, temporal, spatial, and nuanced edges.
- **Event Candidate Extraction (ECE)** —automatically retrieves and summarises real-world events for each claim.
- **Score-Aware Graph Construction** — weights edges with temporal, spatial, semantic, and nuance scores to prioritise high-quality evidence.

2 Related Work

Research on misinformation has been extensively explored, with many studies adopting a binary classification approach. For example, Wang et al. (2020b) propose WeFEND, a reinforcement learning framework designed to filter noisy crowd-sourced reports, addressing the challenge of limited labeled data. While effective for binary fake news detection, WeFEND assumes all claims are either entirely true or entirely false, overlooking borderline or ambiguous cases. Earlier work on multi-class datasets has shown that mixture labels in between true and false are often predicted as hoaxes, mapping mostly to false (Torabi Asr and Taboada, 2018). Not accounting for this gray area can weaken detection, since some online users employ half-truths as propaganda to mislead readers (Hazra and Majumder, 2024). This stresses the importance of considering gray-area class labels. Using the PolitiFact dataset with six labels, the subquestion-based approach (Chen et al., 2022) improved multi-class veracity prediction, yet overall performance remained modest, highlighting the difficulty of distinguishing fine-grained cases such as half-true.

Beyond label design, model architecture also introduces limitations. ICP-BGCN (Hu et al., 2025b) combines tweet content and propagation structure into a graph but ignores external evidence, leaving it prone to echo-chamber bias. FrameTruth (Wang et al., 2024) extracts misleading narrative frames with an LLM, yet its text-only scope overlooks temporal, spatial, and source-level context. CAMOUFLAGE (Bethany et al., 2025) rewrites claims with hedges and ambiguity to evade detectors, but treats hedging solely as adversarial noise rather than an informative signal. More recently, Tang et al. (2025) introduced POLITIFACT-HIDDEN, a 15k-claim dataset annotated with omitted evidence and intent, and proposed TRACER, a framework that models omissions for half-truth detection. Integrated with existing verifiers, TRACER improved Half-True F1 by up to 16 points, underscoring the need to capture hidden context for trustworthy verification.

While several prior studies have explored half-truths, mixture labels, and omitted evidence (Chen et al., 2022; Tang et al., 2025), none have explicitly modelled linguistic nuance as the primary learnable signal for determining borderline claims. Existing approaches often collapse such borderline statements into either ”True” or ”False,” overlooking the linguistic and contextual subtleties that define partial truths. To the best of our knowledge, *MEGA* is the first framework to operationalize *Somewhat True* as an independent, learnable class, treating nuance not as noise but as a structural feature that bridges the gap between traditional binary classification and a more complex real-world claims.

In summary, prior work often relies on binary labels, internal propagation graphs, or text-only framing models, and sometimes treats linguistic nuance as noise. Our framework addresses this by modelling nuance with both a controllable module and a scoring mechanism, while incorporating event evidence and spatio-temporal metadata into the verification process.

3 Methodology

Our proposed framework, MEGA (Multidimensional Event-Guided Analysis), addresses the challenge of classifying borderline misinformation claims by combining real-world evidence, metadata, linguistic tone, and quality signals into a unified graph-based architecture. Our framework has four stages: (1) Event Candidate Extraction (ECE), (2) Nuance Control Module (NCM), (3) Evidence-Quality Assessment Score (EQAS), and (4) MEGA graph construction and classification.

3.1 Event Candidate Extraction(ECE)

The first step is to link each claim c_i (with metadata $m_i = (\text{date}, \text{platform})$) to external real-world evidence. We retrieve an event snippet e_i by generating structured queries using named entities extracted with spaCy (Honnibal et al., 2020), temporal expressions identified via rule-based patterns, and platform-specific keywords.These snippets were submitting to a SearXNG-powered search interface (SearXNG, 2021) for multi-engine lookups. Retrieved passages are embedded with Sentence-BERT (Reimers and Gurevych, 2019), clustered semantically, and summarised into a single factual event snippet e_i.

If search or clustering fails, we return a short "no reliable event context found" note, so downstream steps always receive a clear, interpretable output.

3.2 Nuance Control Module (NCM)

We change tone, not facts. This module manipulates the linguistic tone of event candidates before they are scored and selected, adjusting each event snippet e_i to convey varying levels of clarity, ambiguity, or caution. In this paper, *linguistic tone* refers to surface cues that influence how a statement is read—such as hedges and modality ("may", "could"), conditionality ("if", "unless"), attribution ("according to…"), and contrast markers ("however", "but"). The presence and strength of these cues are referred to as *nuance*.

We apply *linguistic reframing* to modify these nuances without adding or removing factual content. Specifically, we define two transformation mechanisms (Figure 1):

1. **Nuance injection** — introduces hedging/ambiguity (e.g., "reportedly", "suggests", "appears to").
2. **Nuance removal** — eliminates those markers to make the same content more assertive.

Formally, let e_i denote the event snippet retrieved by ECE for claim c_i. The NCM generates two rewrites: an *injected* version e_i^{inj} (adds hedging/ambiguity cues) and a *removed* version e_i^{rem} (strips them). Each experimental variant uses exactly one of these downstream; for brevity, we write

$$e_i^* \in \{\, e_i^{\text{inj}},\, e_i^{\text{rem}} \,\}.$$

We generate e_i^{inj} and e_i^{rem} using Qwen2.5-14B-Instruct hosted locally via Ollama with fixed prompts and parameters to ensure consistency and reproducibility (Bai et al., 2023; Ollama, 2023). Only the event snippet is rewritten; the claim c_i remains unchanged. The resulting pair (c_i, e_i^*) is then used for Evidence-Quality Assessment Score (EQAS) and node-feature construction. This setup lets us directly measure how framing influences classification—especially for *Somewhat True* class.

3.3 Evidence-Quality Assessment Score (EQAS)

For each pair of claim and event snippet (c_i, e_i^*), we compute a four-dimensional score vector $S = \{s_T, s_S, s_M, s_N\}$:

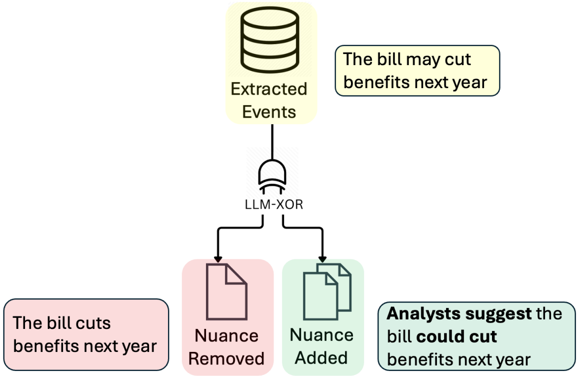


Figure 1: Nuance Control Module (NCM). Given the extracted event snippet, NCM applies *one* rewrite: inject hedging/ambiguity or remove it, producing two alternative snippets used in our variants.

- **Temporal specificity** (s_T) —precision of temporal references in e_i^*, determined via rule-based parsing of explicit dates and scaled to $[0, 1]$.
- **Spatial specificity** (s_S) —granularity of location mentions in e_i^*, mapped by rule-based city/region/country resolution to $[0, 1]$.
- **Semantic similarity** (s_M) —cosine similarity between Sentence-BERT embeddings of c_i and e_i^* (Reimers and Gurevych, 2019).
- **Nuance score** (s_N) —strength of hedging or ambiguity cues in e_i^*, assigned by a locally hosted Qwen2.5-14B-Instruct using a short rubric; computed only when NCM is enabled (Bai et al., 2023).

For claim c_i, we denote $s_i = [s_{T,i}, s_{S,i}, s_{M,i}, s_{N,i}]$, with $s_{N,i}$ omitted when NCM is disabled.

The score set S serves two purposes: (i) pruning edges via adaptive, type-specific thresholds, and (ii) augmenting node features during graph construction, which will be done in the next stage.

3.4 MEGA Graph Construction and Classification

Node Features. Each data point is $d_i = (c_i, m_i, y_i)$, where c_i is the claim text, $m_i = (\text{date}, \text{platform})$ is metadata, and $y_i \in \{0, 1, 2\}$ is the gold label (*Completely False*, *Somewhat True*, *True*). We encode: (1) c_i with SBERT $\to t_i$; (2) m_i into z_i using date buckets and platform one-hots; (3) e_i^* into EQAS $s_i = (s_{T,i}, s_{S,i}, s_{M,i}, s_{N,i})$. The node feature is:

$$x_i = [\, t_i \parallel z_i \parallel \text{enc}(e_i^*) \parallel s_i \,],$$

where $\text{enc}(\cdot)$ is the SBERT embedding of e_i^*. y_i is used only for training and evaluation purposes.

Graph and pruning. We construct a claim—evidence graph $\mathcal{G}$ over all claims, where each node v_i is assigned the feature vector x_i. Edges connect nodes whose claims and associated events are similar in semantic, temporal, or spatial terms, with the corresponding similarity scores stored as edge features.

Adaptive pruning. Using only the training split, we examine the distribution of each edge-score type (semantic, temporal, spatial) and select one cutoff per type (e.g., a chosen percentile). These cutoffs are then fixed and applied unchanged to validation and test splits to avoid leakage. An edge (i, j) is retained if it meets the semantic threshold, or if it satisfies both the temporal and spatial thresholds. We further keep only the top-k most similar neighbours (by semantic score) for each node to prevent any single node from dominating the graph. When the Nuance Control Module (NCM) is active, we increase the thresholds for edges whose endpoints have higher average nuance, $\bar{s}_N = \frac{s_{N,i}+s_{N,j}}{2}$, making the gate stricter when reframing is more ambiguous. This ensures that only well-supported links are preserved in high-nuance contexts.

Classifier. We employ a standard Graph Attention Network (GAT) without architectural modifications (Veličković et al., 2018). The combination of edge-aware construction and adaptive pruning biases the model toward stronger, contextually grounded relationships while reducing noise from weak or misleading connections.

3.5 Dataset and Labelling

We collect fact-checked claims from PolitiFact (2007—2024) (PolitiFact, 2024), including claim text, publish date, platform, and the original veracity label. PolitiFact uses six labels: *Pants on Fire*, *False*, *Mostly False*, *Half True*, *Mostly True*, and *True*.

Model Configuration	Scores	NCM
Claims only	No	No
Claims + metadata	No	No
ECE only	No	No
ECE + EQAS	s_T, s_S, s_M	No
Nuance injected (no EQAS)	s_N	Yes
Nuance removed (no EQAS)	s_N	Yes
sN-only	s_N	Yes
Full MEGA	All	Yes
Contrastive removal	s_N	Yes
ECE Core Isolation	No	No
Positional bias	No	No

Table 1: Feature and edge model configurations used in the experiments

For our experiments, we relabel to three classes to separate outright falsehoods, clear truths, and ambiguous cases:

- **Completely False** —merge *Pants on Fire* + *False*
- **Somewhat True** —merge *Half True* + *Mostly True*
- **True** —keep *True* as-is

We exclude *Mostly False* due to inconsistent annotation patterns and class imbalance in our corpus, which would introduce noise into the three-class distinction we aim to evaluate. The final dataset contains 26,500 labelled claims after cleaning (removing nulls, duplicates, extreme-length outliers, and formatting noise). For a balanced evaluation, we sample 6,000 claims (2,000 per class) with a fixed seed and use this same subset across all experiments.

4 Experiments

4.1 Experimental Setup

We conducted extensive experiments across multiple model configurations as shown in Table 1. All models use two GAT layers with a hidden size of 256 and 8 attention heads (Veličković et al., 2018), with each node linked to its top-7 semantic neighbours. Training uses cross-entropy loss, the AdamW optimiser with a learning rate of 5×10^{-4} (Loshchilov and Hutter, 2019), early stopping after 25 epochs without improvement, and a dropout rate of 0.30. We use sentence-BERT `all-mpnet-base-v2` to encode the text (Reimers and Gurevych, 2019).

Model	F1-Score by Class			Acc.
	T	SW True	CF	
Baseline Models				
Claims only	58	63	60	60
Claims + metadata	64	64	62	63
Real-World Context				
ECE only	72	65	70	69
Nuance Control Variants (no EQAS)				
Nuance injected	**77**	**74**	**73**	**75**
Nuance removed	74	70	72	72
Nuance-injected (EQAS) per dimension				
Nuance Score (s_N)	**78**	**80**	71	**77**
Contextual only	77	74	73	74
Temporal only	**78**	75	70	76
Spatial only	77	76	72	75
Spatial + Contextual	71	74	71	74
Spatial + Temporal	**78**	75	**74**	76
Contextual + Temporal	77	75	71	74
Full MEGA	76	74	73	74

Table 2: Performance metrics across models configurations. Abbreviations: T = True; SW True = Somewhat True; CF = Completely False; Acc. = Accuracy. The values are in percentage, applied for all the subsequent tables

The dataset is split into 70% training, 10% validation, and 20% test sets, stratified by class. We evaluate performance using Accuracy and per-class F1, and analyse confusion matrices to investigate misclassification boundaries, particularly for cases near decision edges (Fawcett, 2006). Unless otherwise stated, all tables report the same 20% test split with identical thresholds and prompts carried over from training.

4.2 Results and Discussion

Impact of External Evidence. Baseline models highlight the difficulty of claim classification without real-world context. The claims-only model reached just 60% accuracy, with weak performance across all labels (Table 1). Adding metadata such as platform and date improved accuracy by 3%, showing limited discriminative value on its own. A larger gain came from external evidence: incorporating ECE snippets raised accuracy to 69%. This supports the premise that linking claims to real-world events provides factual anchors through temporal and spatial cues. However, the model continued to struggle with *Somewhat True*, motivating the need for additional signals.

Nuance injection. The next significant shift occurs when the Nuance Control Module (NCM) introduces hedging and ambiguity into event snippets. Accuracy rises to 75%, with *Somewhat True* F1 improving by +9 points over ECE-only. Gains are also consistent for True and False classes. These improvements indicate that the model is not simply relaxing decision criteria but exploiting tone-related cues that clarify borderline distinctions. In particular, hedging and contrastive phrasing sharpen the boundary between *Somewhat True* and both extremes, showing that linguistic nuance functions as a meaningful signal rather than noise.

Nuance removal. When nuance is removed from the event snippet, the performance still improves compared to the base ECE configuration, with *Somewhat True* rising from 65% to 70% and overall accuracy from 69% to 72%. However, this configuration falls short of the injection gains, with *Somewhat True* reaching 74% and accuracy 75% under injection. This gap suggests that removing linguistic cues helps reduce some confusion but also strips away information that could aid the model in identifying fine-grained distinctions. Without these cues, the boundary between True and *Somewhat True* becomes less defined, and certain borderline cases may be pushed toward the wrong side of the decision threshold. The fact that removal still performs better than base ECE implies that not all nuance is helpful, and in some contexts, tone markers may distract the model from content-based reasoning.

Nuance as Isolated Signal. To examine the effect of linguistic nuance in isolation, the Nuance Score s_N is used as a probe in two settings: using only s_N, and applying the same score to versions where nuanced phrasing has been removed. Using only s_N yields the highest overall accuracy at 77% and the strongest *Somewhat True* F1 at 80%, surpassing the Full MEGA configuration, which achieves 74% accuracy. When s_N is applied to the stripped versions, performance declines in proportion to the amount of nuance removed, indicating that s_N captures the influence of linguistic tone rather than memorising content. The comparative results are shown in Table 3. The values for nuance injection and removal differ from those in the previous table because this experiment measures the effect of nuance alone, without other cues. This indicates that s_N alone is a strong proxy for linguistic tone.

Configuration	T	SW True	CF	Acc.
Nuance injection	**78**	**80**	71	**77**
Nuance removal	78	72	**72**	75
Contrastive removal	72	64	70	69

Table 3: Nuance Score s_N variants.

Score Config.	T	SW True	CF	Acc.
Contextual only	75	66	69	70
Temporal only	73	**77**	70	70
Spatial only	72	63	70	68
Spatial + Contextual	71	65	**71**	69
Spatial + Temporal	75	67	**71**	**71**
Contextual + Temporal	**76**	65	**71**	70
All combined	73	66	70	70

Table 4: EQAS applied to base ECE.

The Effect of Evidence-Quality Assessment Score (EQAS) Module. Applying EQAS on top of the base ECE produces only modest changes in performance (Table 4). Overall accuracy ranges from 68% to 71%, with the highest at 71% for the Spatial + Temporal configuration, a gain of two points over ECE-only at 69%. The Temporal-only setting pushes the *Somewhat True* F1 to 77% but does not raise overall accuracy beyond 70%. Other configurations mostly exchange small gains between classes without a consistent advantage. While these results show that EQAS adds useful signal, its contribution is secondary to the larger improvements achieved through nuance.

When nuance is reduced—either by removing all nuanced phrasing or only contrastive cues—EQAS still provides measurable but modest gains (Tables 5 and 6). Temporal and spatial scores occasionally lift accuracy by up to two points over the base setting, with Temporal-only and Spatial-only configurations performing best in their respective contexts. This shows that EQAS retains value even without nuanced language, but its effect is smaller and less consistent than when nuance is preserved (see Table 2).

Full MEGA Configuration. Full MEGA is the complete configuration of our framework, combining the ECE evidence snippet

Score Config.	T	SW True	CF	Acc.
Contextual only	72	73	72	72
Temporal only	**77**	**75**	**74**	**75**
Spatial only	74	72	72	73
Spatial + Contextual	73	74	72	73
Spatial + Temporal	77	**75**	73	**75**
Contextual + Temporal	70	73	70	71
All combined	74	74	71	74

Table 5: EQAS with all nuance removed.

Score Config.	T	SW True	CF	Acc.
Contextual only	72	66	70	69
Temporal only	74	66	**71**	70
Spatial only	**76**	66	70	**71**
Spatial + Contextual	73	66	70	70
Spatial + Temporal	73	66	69	69
Contextual + Temporal	75	**69**	70	**71**

Table 6: EQAS after contrastive removal.

e_i, an NCM rewrite e_i^*, and all EQAS scores $S = \{s_T, s_S, s_M, s_N\}$, which are encoded in the node features and also used as edge signals in the graph. As shown in Table 2, this configuration delivers strong and balanced performance across classes, although it is not the top performer for *Somewhat True*, where the nuance-injected model without EQAS achieves slightly higher F1 and accuracy. We evaluated both configurations on unseen claims, keeping all thresholds, hyper-parameters, and model settings fixed. Both maintain an F1 of 75% on *Somewhat True*, indicating that the nuance signal generalises beyond the training distribution. Full MEGA achieves the highest overall accuracy in this setting (77% vs. 76% for the nuance-injected variant) by combining temporal and spatial gating with semantic evidence, which slightly reduces off-class errors (Table 7).

We therefore regard Full MEGA as the comprehensive, stability-oriented configuration, while the nuance-injected variant without EQAS remains the most effective for borderline detection.

Model Variant	T	SW True	CF	Acc.
Nuance-injected ECE	77	**75**	75	76
Full MEGA	**78**	**75**	**76**	**77**

Table 7: Generalisation performance on unseen claims

Nuance Config.	T	SW True	CF	Acc.
Original (front-loaded)	77	74	73	75
Mid-loaded	75	72	73	73
Back-loaded	74	77	73	75

Table 8: Impact of shifting nuance position within a claim.

4.3 Diagnostics: Examining Model Behaviour

We conducted three controlled experiments to disentangle the contribution of linguistic nuance from other model cues: (1) *Positional Bias* —hedging cues (e.g., "may cause") were moved to different positions in the sentence (front, middle, end) to test whether their location influences predictions. (2) *Contrastive Framing* —rhetorical pivots such as "however" and "although" were removed to evaluate reliance on explicit discourse contrast. (3) *Core Isolation* —each event was reduced to its factual core, removing all hedging, elaboration, and contextual detail, to assess how structural simplification affects classification.

Structural dependency via positional bias. The positional bias test examined whether the location of nuance changes the model's decision-making. As shown in Table 8, shifting hedging cues had minimal effect, with only a 2% drop in accuracy for mid-position placement. This suggests the model's detection of nuance is not tied to its syntactic location but rather to its lexical and semantic presence in the sentence. Performance stability across positions indicates that nuanced phrasing is treated as a content-level signal rather than a positional signal.

Contrastive removal (rhetorical pivots). The contrastive framing test evaluated the impact of removing explicit discourse markers that signal rhetorical shifts. Compared to the Full MEGA baseline, removing cues such as "however" and "although" reduced accuracy (Table 9), with the largest relative drop in *Somewhat True* performance. These pivots typically mark stance changes or qualifications, making them especially informative for detecting borderline or mixed-truth claims. Their removal reduces the model's ability to recognise such shifts, confirming that contrastive phrasing acts as a nuance-like signal in classification.

Model	F1-Score by Class			Acc.
	T	SW True	CF	
Nuance-Focused Baselines				
Nuance injection	**77**	74	**73**	**75**
Nuance Score (s_N)	**78**	**80**	71	**77**
Full MEGA	76	74	73	74
Structural Diagnostics				
Contrastive removal	76	68	72	72
ECE Core Isolation	**79**	**83**	**76**	**80**

Table 9: Comparison of nuance-focused models and structural diagnostic variants

Core isolation (higher-accuracy pitfall). Finally, we investigated the effect of stripping away all structural tone. The Core Isolation variant (which reduces events to bare factual statements without hedging or contextual detail) yielded the highest raw accuracy among non-EQAS settings (Table 9), but this created a problematic trade-off. As shown in the confusion matrices (Figures 2–3), predictions skewed toward extreme labels, particularly *Completely False*. Counts rose from 259 in the nuance-injected variant to 307 under Core Isolation, with "True" → "False" errors increasing from 12 to 18, and *Somewhat True* → "False" from 46 to 54. Thus, accuracy gains came at the cost of misclassifying borderline cases, indicating sharper but less calibrated decision boundaries.

Interpreting the results. Event grounding (ECE) was necessary but not sufficient—linking claims to real-world events provided the first performance lift. The decisive change came from linguistic nuance: injecting hedging and conditional cues prevented the collapse of borderline cases into extremes, allowing the model to treat nuance as a distinct, learnable signal rather than noise. In contrast, Core Isolation simplified the problem rather than solving it, improving accuracy for the wrong reason by inflating binary decisions.

Nuance therefore acts as a dual-role structural signal. As text, it consistently stabilises *Somewhat True* predictions; as a graph feature, it retains influence via the nuance score, providing a direct input for model reasoning. These effects are position-independent, and contrastive phrasing behaves similarly to nuance, broadening the operational definition of nuanced language. EQAS complements this by anchoring decisions to temporal, spatial, and semantic context, but its impact is secondary when strong tone cues are present. Overall, the most robust configuration is ECE + Nuance Injection (no EQAS), which preserves calibration on borderline content while still generalising effectively to unseen claims.

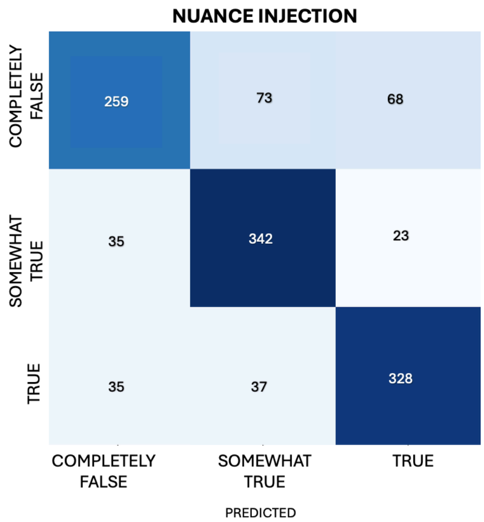


Figure 2: Confusion matrix for ECE with nuance injection.

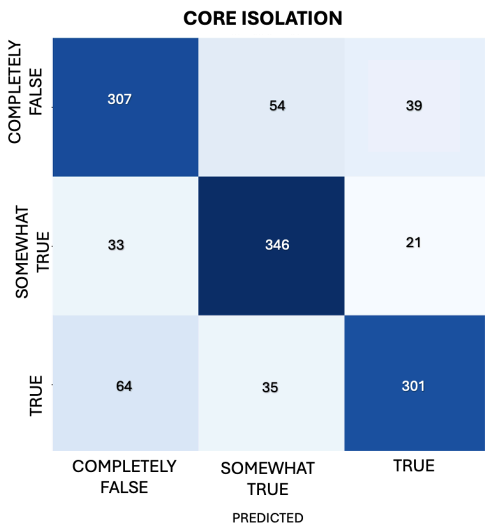


Figure 3: Confusion matrix for the ECE core isolation experiment.

5 Conclusion

Nuance stands out as the signal that defines our approach to misinformation detection. Real-world event grounding provides evidential anchoring, but it is the modelling of tone—hedging, conditionality, and contrast—that consistently enables accurate recognition of partial truths. This effect holds regardless of where cues appear, showing that their strength comes from presence, not position. Other signals, like temporal, spatial, and semantic scores, add stability but do not replace the interpretive weight of nuance. By embedding this signal into both the evidence and the graph, we show that subtle language patterns are not noise, but essential, learnable features for distinguishing misinformation with precision.

Limitations and Future Works

Our framework adopts a relatively simple architecture that combines Sentence-BERT embeddings with a Graph Attention Network, allowing us to isolate and highlight the effects of linguistic nuance. This design effectively captures the contribution of tone and event context; however, its simplicity may constrain the model's expressive capacity and ultimate performance ceiling. Consequently, the full potential of nuanced language understanding within state-of-the-art fact-verification architectures, which incorporate richer contextual modeling or explicit propagation dynamics, remains an open area for further exploration.

Recent fact-verification models use dense passage retrieval (Thorne et al., 2018), fine-tuned transformers trained on large-scale verification datasets (Schuster et al., 2019), or heterogeneous graphs that capture social propagation patterns (Hu et al., 2025b). Such architectures may already capture hedging and tonal variation through large-scale pre-training or by integrating evidence from multiple sources. However, it remains uncertain whether these implicit signals achieve the same interpretive precision as explicit nuance modeling. In other words, while advanced models may recognize linguistic uncertainty to some extent, they may not yet distinguish how specific tone markers influence veracity judgments.

Future work could therefore explore integrating the ECE and NCM modules into more advanced architectures would yield diminishing returns or, conversely, reveal complementary effects—and how stronger baselines might interact with nuance-aware modelling to either enhance or reduce their overall impact.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Ma, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Xu, Zhicu Yang, Zhenru Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Tianzhu Zhang, Bowen Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Mazal Bethany, Nishant Vishwamitra, Cho-Yu Jason Chiang, and Peyman Najafirad. 2025. Camouflage: Exploiting misinformation detection systems through llm-driven adversarial claim transformation. arXiv preprint arXiv:2505.01900.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shaobo Cui, Junyou Li, Luca Mouchel, Yiyang Feng, and Boi Faltings. 2025. Nuance matters: Probing epistemic consistency in causal reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23715–23723.

T. Fawcett. 2006. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Sanchaita Hazra and Bodhisattwa Prasad Majumder. 2024. To tell the truth: Language of deception and language models. 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 8506–8520, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Jie Hu, Mei Yang, Bingbing Tang, and Jianjun Hu. 2025a. Integrating message content and propagation path for enhanced false information detection using bidirectional graph convolutional neural networks. *Applied Sciences*, 15(7):3457.

Jie Hu, Mei Yang, Bingbing Tang, and Jianjun Hu. 2025b. Integrating message content and propagation path for enhanced false information detection using bidirectional graph convolutional neural networks. *Applied Sciences*, 15(7):3457.

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Ollama. 2023. Run large language models locally. https://github.com/ollama/ollama.

PolitiFact. 2024. Politifact fact-check database. https://www.politifact.com/. Accessed 2025-07-10.

H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2931–2937.

N. Reimers and I. Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992.

Victoria L. Rubin. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144, Rochester, New York. Association for Computational Linguistics.

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel R Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

SearXNG. 2021. A privacy-respecting metasearch engine. https://github.com/searxng/searxng.

K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, 10(3):1–42.

Yixuan Tang, Jincheng Wang, and Anthony K. H. Tung. 2025. The missing parts: Augmenting fact verification with half-truth detection.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Fatemeh Torabi Asr and Maite Taboada. 2018. The data challenge in misinformation detection: Source reputation vs. content veracity. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 10–15, Brussels, Belgium. Association for Computational Linguistics.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations (ICLR)*.

Guan Wang, Rebecca Frederick, Boshra Talebi Haghighi, B. L. William Wong, Verica Rupar, Weihua Li, and Quan Bai. 2024. Frametruth: A frame-based model utilizing large language models for misinformation detection. In *Proceedings of ACIIDS 2024 (LNAI 14795)*, pages 135–146.

Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, and J. Gao. 2020a. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6359–6369.

Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020b. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*, pages 516–523, New York, NY, USA.

A Preliminary Study of RAG for Taiwanese Historical Archives

Claire Lin[1*], Bo-Han Feng[2*], Xuanjun Chen[3*], Te-Lun Yang[4]
Hung-yi Lee[3], Jyh-Shing Roger Jang[2,4]
[1]Department of Information Management, National Taiwan University
[2]Department of Computer Science and Information Engineering, National Taiwan University
[3]Graduate Institute of Communication Engineering, National Taiwan University
[4]Graduate Institute of Networking and Multimedia, National Taiwan University
{b10705004, b10902031, d12942018, d12944007}@ntu.edu.tw
hungyilee@ntu.edu.tw, jang@mirlab.org

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a promising approach for knowledge-intensive tasks. However, few studies have examined RAG for Taiwanese Historical Archives. In this paper, we present an initial study of a RAG pipeline applied to two historical Traditional Chinese datasets, Fort Zeelandia and the Taiwan Provincial Council Gazette, along with their corresponding open-ended query sets. We systematically investigate the effects of query characteristics and metadata integration strategies on retrieval quality, answer generation, and the performance of the overall system. The results show that early-stage metadata integration enhances both retrieval and answer accuracy while also revealing persistent challenges for RAG systems, including hallucinations during generation and difficulties in handling temporal or multi-hop historical queries.

Keywords: Retrieval-Augmented Generation, Humanities Data, Large Language Model

1 Introduction

Recent advances in large language models have substantially improved open-domain question answering and knowledge-intensive tasks. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), which combines document retrieval with text generation, has shown promise in mitigating hallucination and improving factuality. Prior research has primarily focused on English (Bajaj et al., 2018; Kwiatkowski et al., 2019; Yang et al., 2024) or Simplified Chinese datasets (Lyu et al., 2024; Li et al., 2024a) and general-purpose domains such as Wikipedia or web-collected questions.

However, much less attention has been given to RAG performance on underrepresented languages and culturally specific corpora, particularly in the humanities. Historical contexts in Traditional Chinese pose unique challenges, including unstructured documents, time-sensitive content, and linguistic differences between queries and archival sources. These factors complicate both retrieval and generation, making it unclear how well current RAG systems handle such materials.

To address this gap, we propose two Taiwanese historical datasets, Fort Zeelandia and Taiwan Provincial Council Gazette (TPCG), along with their associated query sets, as case studies for historical open-ended question answering. The datasets are annotated with query-level and document-level metadata, enabling fine-grained experiments on how query types and metadata integration strategies affect RAG performance. Through systematic evaluation across multiple retrieval methods and query characteristics, we demonstrate that early-stage metadata integration substantially improves system effectiveness. Furthermore, our findings reveal persistent challenges: hallucinations remain a recurring issue during generation, and questions involving temporal reasoning exhibit notable difficulty. Furthermore, our analysis of retrieval performance reveals that multi-hop and time-sensitive queries tend to yield lower recall, whereas early-stage metadata integration consistently delivers the strongest overall retrieval effectiveness.

2 Related Work

RAG Lewis et al. (2020) improves language model performance on knowledge-intensive tasks by incorporating relevant external information during generation. By grounding outputs in retrieved evidence, RAG reduces hallucinations when models encounter unfamiliar topics and alleviates the substantial cost of continuously retraining models to incorporate new knowledge.

[*]Equal contribution.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 45–62
November 20-22, 2025 ©2025 Association for Computational Linguistics

Dataset	Language	Humanities	Query-Passage Pairs	Metadata
MS MARCO	English	✗	✓	Limited
Natural Questions	English	✗	✓	✗
MMLU	English	✓	✗	✗
CMMLU	Simplified Chinese	✓	✗	✗
Fort Zeelandia Query Set (Our)	Traditional Chinese	✓	✓	✓
TPCG Query Set (Our)	Traditional Chinese	✓	✓	✓

Table 1: Comparison of datasets by language, domain knowledge, structure, and metadata. Fort Zeelandia and TPCG Query sets stand out for their rich metadata and grounding in historical or contextual knowledge.

Early benchmarks of RAG mainly relied on general-purpose datasets such as MS MARCO (Bajaj et al., 2018) and Natural Questions (Kwiatkowski et al., 2019). More recently, researchers have introduced domain-specific datasets in areas including biomedicine (Xiong et al., 2024; Li et al., 2024b; He et al., 2025), law (Pipitone and Alami, 2024; Zheng et al., 2025; Wahidur et al., 2025), and non-English languages such as Traditional Chinese (Yang et al., 2025). However, RAG applications in the humanities are underexplored, particularly for Taiwanese historical materials.

Table 1 compares the key differences of existing benchmarks with the query sets from our newly introduced Fort Zeelandia and TPCG datasets. Firstly, in terms of humanities coverage, MS MARCO and Natural Questions primarily target general-purpose or factual QA and contain little to no humanities material, whereas MMLU (Hendrycks et al., 2021) and CMMLU (Li et al., 2024a) include partial coverage through their broader topical scope. By contrast, our Fort Zeelandia and TPCG query sets are explicitly designed around humanities data, with a particular emphasis on historical materials. Secondly, with respect to query–passage alignment, MS MARCO and Natural Questions are constructed around paired queries and passages, a design we also adopt for Fort Zeelandia and TPCG query sets to support retrieval-based evaluation. MMLU and CMMLU, in contrast, rely on multiple-choice formats. Finally, in terms of metadata, our proposed datasets provide rich query- and document-level annotations, enabling more fine-grained retrieval experiments and analysis than existing resources.

3 Dataset

We introduce two Traditional Chinese datasets from Taiwanese historical archives: Fort Zeelandia and Taiwan Provincial Council Gazette (TPCG). We refer to the associated queries as the Fort Zeelandia Query Set and the TPCG Query Set, and to Fort Zeelandia and TPCG themselves as the document datasets in this paper.

3.1 Fort Zeelandia

Entity	Single-hop	Multi-hop	Total
Event	32	18	50
Item	14	2	16
People	19	4	23
Place	16	6	22
Time	19	4	23
Multi-entity	0	39	39
Total	100	73	173

Table 2: Fort Zeelandia Query Set Entity Focus Distribution across Question Complexity

This dataset is constructed from historical diaries[1] documenting Dutch colonization of Taiwan in the 17th century. We collected 5,443 passages and collaborated with students from the Department of History, who created 173 queries and annotated the relevant passages for each query.

Query-level Metadata. Each QA pair is annotated with query-level metadata, including:

- **Question complexity**: Single-hop or multi-hop question. A multi-hop question requires combining information from multiple passagesto determine the answer, whereas a single-hop question can be answered using just one passage.
- **Entity focus**: Whether the question centers on a person, item, time, event, or location.

An example from the Fort Zeelandia dataset is demonstrated in Appendix A.1.

3.2 Taiwan Provincial Council Gazette

The TPCG dataset comprises official meeting records from the Taiwan Provincial Council As-

[1] https://taco.ith.sinica.edu.tw/tdk/

sembly [2], spanning the mid to late 20th century, totaling 228,135 documents. To build the question answering benchmark, history students manually crafted 56 question-passage pairs based on selected gazette excerpts. The resulting dataset captures realistic information needs and research scenarios commonly encountered in historical inquiry.

Document-level Metadata. TPCG is characterized with well-defined document-level metadata, enabling experiments on how structured context can be used to improve system performance. Each document is associated with:

- **Time/Event Information**: Includes time information such as the start and end dates, volume and published date.
- **Person/Organization Information**: Covers participating members, agencies, decree, presiding officials and president at that time.
- **Content/Document Information**: Includes document title, abstract, content type, category, subject, keywords, attachments, references, and remarks.

An example from the TPCG dataset is demonstrated in Appendix A.2.

4 Methods

The RAG pipeline in Figure 1 comprises four stages: Input, Retrieval, Generation, and Evaluation. Throughout the pipeline, we (a) construct datasets and annotate query–passage pairs, (b) retrieve candidate passages using lexical, dense, and hybrid methods with optional metadata integration and reranking, (c) prompt a generator LLM with the query, retrieved passages, and metadata to generate an answer, and (d) assess answer quality with an LLM-as-judge protocol.

4.1 Input

The input stage in Figure 1 (a) covers data acquisition and annotation. We first crawl and normalize raw materials into document collections for Fort Zeelandia and TPCG datasets. Domain experts (Taiwanese history students) then author queries and annotate the associated gold passages, yielding high-quality query–passage pairs for RAG experimentation. To enable controlled analysis, we further annotate (i) question complexity (single-hop vs. multi-hop) and entity focus (people, event, time, place, item, or multi-entity) for Fort Zeelandia, and (ii) document-level metadata for TPCG, grouped into Time/Event, Person/Organization, and Document/Content categories.

[2]https://drtpa.th.gov.tw/index.php?act=Archive

4.2 Retrieval

Given a user query, the retrieval stage in Figure 1 (b) identifies a small set of passages most likely to support grounded answer generation. This stage is essential in a RAG pipeline because it (i) grounds the generator in verifiable evidence to reduce hallucinations, (ii) filters a large corpus into a compact candidate set that fits the context window, and (iii) adapts to lexical, semantic information, and structured metadata in Fort Zeelandia and TPCG. The stage comprises two parts: retrieval models (sparse, dense, hybrid) that score query–passage relevance, and retrieval strategies that optionally use document-level metadata and a second-stage reranker. Together, these components return top-k passages for the generation stage.

4.2.1 Retrieval Models

We instantiate three families of retrieval models:

Sparse retrieval. We adopt BM25 (Robertson and Zaragoza, 2009), which retrieves documents based on term-matching style term-frequency and inverse document frequency (TF-IDF) weighting (Salton and Buckley, 1987), together with sparse embeddings derived from a BGE-M3-based model (Chen et al., 2024).

Dense retrieval. A BGE-M3–based dense encoder maps queries and passages into a shared embedding space for semantic matching, which is helpful when relevant evidence is phrased differently from the query.

Hybrid retrieval. To leverage both lexical and semantic signals, we fuse the sparse and dense ranked lists using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009):

$$\mathrm{RRF}(d) = \sum_{i=1}^{n} \frac{1}{k + r_i(d)} \tag{1}$$

where d is the document, n is the number of ranked lists, $r_i(d)$ is the rank of document d in the i-th ranked list, and k is a constant that dampens the contribution of the lower-ranked documents.

4.2.2 Retrieval Strategies

Beyond first-stage retrieval, we integrate document-level metadata and a second-stage reranker to im-

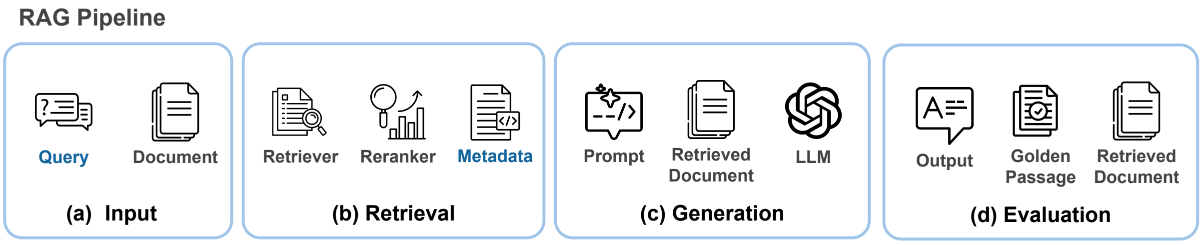


Figure 1: Overview of RAG pipeline and components in each stage. The two highlighted elements: **Query** and **Metadata** are the key factors that impact RAG system performance we focused on in this paper. The details of these factors are discussed in Section 3.1 and Section 3.2, respectively. Section 6.2 and Section 6.3 elaborates how these factors impact retrieval and generation performance.

prove ranking. Metadata in TPCG is grouped into Time/Event, Person/Organization, and Content/Document fields; these fields capture signals (e.g., publication dates, presiding officials, content categories) that are often only weakly expressed in raw text but crucial for precise matching in civic or historical domains. We adopt four strategies, illustrated in Figure 2.

Baseline Retrieval. Retrieve using only the query and original document text without metadata. This provides a clean reference that relies purely on text similarity.

Metadata-Augmented Retrieval. Append selected metadata fields to each document chunk before embedding, treating metadata as part of the content. This allows the retriever to encode, for instance, dates, roles, or categories directly into passage representations so they influence similarity at retrieval time. The retriever returns top-k passages given the embeddings of query and metadata-augmented document chunks.

Metadata-Only Reranking. Incorporate metadata at the reranking stage rather than directly appended to the documents. We first retrieve the top-100 candidate passages using the original documents. Then, compute the similarity between the query and the available document-level metadata of each candidate passage. The passages are reranked based on this similarity score, and the final top-k passages are returned for generation.

Metadata-Augmented Reranking. Append metadata to the original document text before computing similarity for reranking. After retrieving the candidate passages, we concatenate each document's metadata with its original content, and then measure the similarity between this augmented text and the query to rerank the candidates. The top-k passages are returned for generation.

By comparing these strategies, we aim to quantify the contribution of metadata at both embedding and reranking stages, and to better understand how different integration points influence retrieval effectiveness for historical information retrieval.

4.3 Generation

We use GPT-4o (OpenAI et al., 2024) to produce answers conditioned on the retrieved passages. The goal is to leverage an LLM to aggregate information dispersed across multiple relevant passages into a fluent natural-language response.

At inference time, each query is paired with the top-5 retrieved passages and any available metadata, which together serve as the external knowledge context for generation. The model is instructed to ground its answer strictly in the provided materials and to avoid introducing external knowledge not mentioned in the documents. When multiple passages support the same fact, the model is encouraged to prioritize such corroborated information. If none of the provided materials is relevant to the query, the model is instructed to respond with “I don’t know”. The full generation prompt is detailed in Appendix A.3.

4.4 Evaluation

We evaluate both retrieval performance and end-to-end RAG quality. For retrieval evaluation, we report Recall@k, which measures the ratio of relevant passages that appear in the top-k retrieved results for each query:

$$\text{Recall@k} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\text{Relevant}_i \in \text{Top-}k\right) \quad (2)$$

where N is the number of relevant passages for the query, $\mathbb{I}(\cdot)$ is the indicator function, Relevant_i is

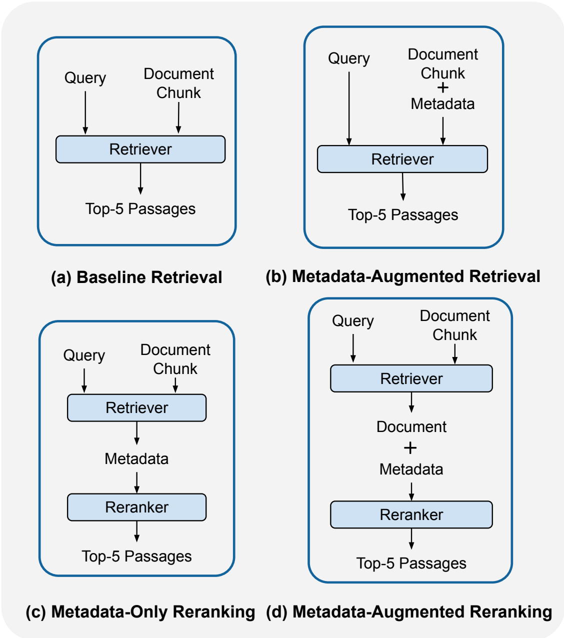


Figure 2: Overview of four retrieval strategies with different metadata integration stages explored in this work. (a) **Baseline Retrieval** retrieves top passages using only the query and document content. (b) **Metadata-Augmented Retrieval** integrates metadata into the document representation during retrieval. (c) **Metadata-Only Reranking** uses only metadata during the reranking stage after initial retrieval. (d) **Metadata-Augmented Reranking** incorporates both document content and metadata in the reranking stage.

the ith relevant passage, and Top-k denotes the top-k retrieved passages. The average Recall@k across all queries yields the overall retrieval performance.

For generation quality, we employ Gemini-2.5-Pro (Comanici et al., 2025) as an evaluator following (Chiang and Lee, 2023). The evaluator is given the golden passage, the retrieved top-5 passages, and the answer from GPT-4o. The complete evaluation prompt is provided in Appendix A.4. It consists of three scoring dimensions: groundedness, relevance, and hallucination.

Groundedness. Assesses whether the generated answer correctly incorporates information directly supported by the golden passage. Each distinct atomic fact from the golden passage that appears correctly in the answer receives one point.

Relevance. Evaluates whether the answer includes additional information present in other retrieved passages consistent with the golden passage. Each relevant atomic fact receives one point.

Hallucination. Penalizes content that is unsupported or irrelevant. For each hallucinated statement or extraneous detail that is neither aligned with the golden passage nor substantiated by the retrieved passages, one point is deducted.

5 Experimental Setup

In our experiments, each document is segmented into chunks of 512 tokens with an overlap of 128 tokens to preserve contextual continuity. For direct retrieval methods, where reranking is not applied, both BM25 and BGE-M3-based approaches are configured to return the top 5 most relevant passages (i.e., top-k = 5). The hybrid method independently retrieves 5 passages using both the sparse and dense retrievers, then combines the two ranked lists using RRF, setting $k = 60$, to produce the final top-5 results. For experiments involving reranking, we first retrieve the top-100 candidate passages and then apply reranking using BGE-reranker (Xiao et al., 2023) to select the final top-5 results. In the reranking scenario, the hybrid approach similarly retrieves 100 passages from each retriever before merging and reranking. We do not perform any retriever and reranker tuning; all retrievers and reranker are used off-the-shelf.

For Fort Zeelandia and its query set, we use passages retrieved by a hybrid retriever with baseline retrieval. For TPCG and the associated query set, we fix the retriever to BM25 and evaluate the impact of different metadata integration stages and types on answer quality. GPT-4o is used to generate answers with the retrieved passages, and Gemini 2.5 Pro is used as an independent evaluator.

6 Results

Figure 1 illustrates the RAG pipeline and its key components at each stage. To evaluate the applicability of the RAG system on historical materials, we conduct experiments using Fort Zeelandia, TPCG, and their query sets. Our study examines how different retrieval strategies, query characteristics, and metadata integration approaches affect overall system performance. The evaluation focuses on multiple dimensions, including the ability to leverage accurate context and the extent of hallucinations.

6.1 Overall RAG Results

Tables 3 and 4 show the overall RAG results on the Fort Zeelandia and TPCG datasets. In Table 4, Metadata-Augmented Retrieval with early Document/Content metadata achieves the highest groundedness, with a significant increase of

Question Type	Subcategory	Groundedness ↑	Relevance ↑	Hallucination ↑
All Questions	-	2.9769	1.0578	-0.6821
Question Complexity	Single-hop	2.8600 (-0.1169)	0.8700 (-0.1878)	-0.5600 (+0.1221)
	Multi-hop	3.1370 (+0.1601)	1.3151 (+0.2573)	-0.8493 (-0.1672)
Entity Focus	People	3.2174 (+0.2405)	1.0870 (+0.0292)	-0.5217 (+0.1604)
	Event	3.4600 (+0.4831)	1.2200 (+0.1622)	-0.5800 (+0.1021)
	Time	1.3478 (-1.6291)	0.4783 (-0.5795)	-0.9565 (-0.2744)
	Place	1.8636 (-1.1133)	1.2273 (+0.1695)	-0.7727 (-0.0906)
	Item	2.5625 (-0.4144)	0.1875 (-0.8703)	-0.5625 (+0.1196)
	Multi-entity	3.9744 (+0.9975)	1.4359 (+0.3781)	-0.7436 (-0.0615)
All Questions (Oracle)	-	4.4104	0.2312	-0.2601

Table 3: RAG evaluation by Query Type on the Fort Zeelandia dataset. The table reports average scores for three evaluation metrics: **Groundedness** (incorporates gold passage information), **Relevance** (integrates relevant passages information), and **Hallucination** (including irrelevant information). For all three metrics, higher values indicate better performance. Since Hallucination scores are negative, a value closer to zero reflects fewer hallucinations. All values are compared against the ”All Questions” row. Colored deltas in parentheses indicate the difference from the average: green for improvement and red for decline. The Oracle row denotes the upper bound of the LLM’s performance when directly given the gold passages. An evaluation example can be found in Appendix A.5.

Integration Stage	Metadata Type	Groundedness ↑	Relevance ↑	Hallucination ↑
Baseline	-	0.7321	0.8571	-0.2500
Metadata-Augmented Retrieval	Time/Event	1.0893 (+0.3572)	1.0000 (+0.1429)	-0.2857 (-0.0357)
	Person/Organization	1.1786 (+0.4465)	0.7321 (-0.1250)	-0.2679 (-0.0179)
	Document/Content	2.1429 (+1.4108)	1.2500 (+0.3929)	-0.3214 (-0.0714)
Metadata-Only Reranking	Time/Event	0.3393 (-0.3928)	1.0000 (+0.1429)	-0.4821 (-0.2321)
	Person/Organization	0.5714 (-0.1607)	0.6071 (-0.2500)	-0.2857 (-0.0357)
	Document/Content	1.5893 (+0.8572)	1.8571 (+1.0000)	-0.3393 (-0.0893)
Metadata-Augmented Reranking	Time/Event	1.2679 (+0.5358)	1.0357 (+0.1786)	-0.6250 (-0.3750)
	Person/Organization	0.9821 (+0.2500)	1.1071 (+0.2500)	-0.6250 (-0.3750)
	Document/Content	1.3750 (+0.6429)	1.0536 (+0.1965)	-0.5357 (-0.2857)
Oracle	-	3.6964	0.0179	-0.0714

Table 4: RAG evaluation by Metadata Integration Strategies on the TPCG dataset. The table reports average scores across the three evaluation metrics. All rows are compared to the Baseline Retrieval, values in the parentheses indicate the improvement or decline. The Oracle row denotes the upper bound of the LLM’s performance when directly given the gold passages. Two evaluation examples can be found in Appendix A.6.

1.4108 over the baseline. Appendix A.7 details the significance tests for various retrieval methods. Performance also varies by query type: event-related queries benefit most, with groundedness up 0.4831, relevance by 0.1622, and hallucinations reduced 0.1021. These findings indicate that RAG effectiveness depends on query characteristics and is strengthened by metadata-augmented retrieval, though hallucinations persist even with oracle passages, highlighting a key limitation.

6.2 RAG Results

This section takes a deeper dive into two key factors that critically influence RAG performance at the Input and Retrieval stages: query type and use of document-level metadata. Specifically, we analyze how different query types affect accuracy, relevance, and hallucination. Additionally, we examine the impact of metadata integration at different stages of retrieval and reranking, considering multiple metadata types. This analysis highlights which combinations of query characteristics and metadata strategies yield the most reliable and accurate outputs for historical open-ended QA tasks.

1) Different Query Types Table 3 illustrates RAG performance across query types. Multi-hop and Multi-entity questions are high-risk: when successful, groundedness increases by 0.1601 and 0.9975, and relevance by 0.2573 and 0.3781, but hallucination worsens by -0.1672 and -0.0615, highlighting a trade-off between complexity and reliability. People- and event-focused queries are

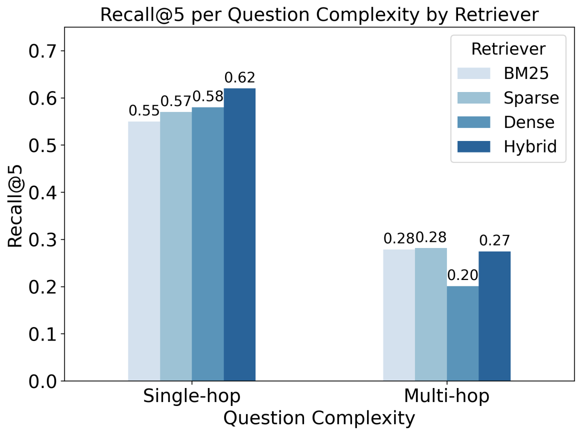


Figure 3: Fort Zeelandia Dataset Recall@5 per Question Complexity by Retriever

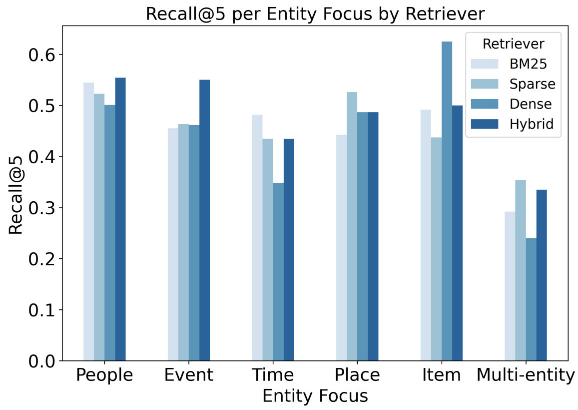


Figure 4: Fort Zeelandia Dataset Recall@5 per Entity Focus by Retriever

more stable, achieving gains in groundedness and relevance with lower hallucination. Time-focused queries are the most challenging, with groundedness and relevance decreasing by 1.6291 and 0.5795, alongside worse hallucination, indicating that temporal reasoning remains a key bottleneck.

2) Different Metadata Integration Strategies Table 4 presents the evaluation scores across three dimensions for the open-ended question answering task, focusing on the key factor Metadata, using TPCG and its query set.

Overall, Metadata-Augmented Retrieval proves the most reliable approach, improving groundedness and relevance with minimal worsening in hallucination. By contrast, reranking strategies show mixed results: Metadata-Only Reranking underperforms the baseline, while Metadata-Augmented Reranking achieves gains in retrieval quality but at the cost of greater hallucination, making it less stable. Across all strategies, Document/Content metadata emerges as the most effective type, underscoring its importance for enhancing the system.

6.3 Ablation Study of Retrieval Results

In this section, we take a closer look at the Retrieval stage of the RAG pipeline. Since RAG fundamentally relies on retrieved documents as the foundation for generating answers, understanding retrieval effectiveness is critical to interpreting overall system performance. By analyzing how different retrieval strategies, query types, and metadata integration methods influence the quality of retrieved context, we can better identify the factors that drive successes and failures in retrieval.

1) Retrieval with Query-level Metadata We investigate the impact of query types on retrieval performance using query-level metadata, focusing on query complexity and entity focus.

Different Question Complexity. To gain deeper insight into RAG performance across varying query complexity, we further analyze the retrieval results on the Fort Zeelandia dataset. Figure 3 presents Recall@5 scores comparing single-hop and multi-hop questions across different retrievers. For single-hop questions, Recall@5 scores are roughly twice as high as for multi-hop questions, corresponding to a lower tendency for hallucination. In contrast, retrievers achieve Recall@5 of at most only 0.28 for multi-hop queries, increasing the likelihood of hallucinated responses.

Notably, despite the lower recall, multi-hop and multi-entity questions still achieve higher groundedness and relevance, suggesting that the LLM is capable of performing multi-step reasoning when appropriate context is provided.

Different Entity Focus. We analyze retrieval performance across different entity focuses to better understand its impact on RAG outcomes. Figure 4 presents Recall@5 scores for People, Event, Time, Place, Item, and Multi-entity questions. For the hybrid retriever used in the RAG pipeline for Fort Zeelandia, performance is notably higher for People- and Event-focused questions, with Recall@5 around 0.55, corresponding to better-controlled hallucination. In contrast, Time- and Multi-entity questions exhibit lower retrieval performance, with Recall@5 of 0.43 and 0.33, respectively, which aligns with increased hallucination.

Considering both RAG scores and retrieval results, we find that although retrieval for Time-focused questions is slightly better than for Multi-

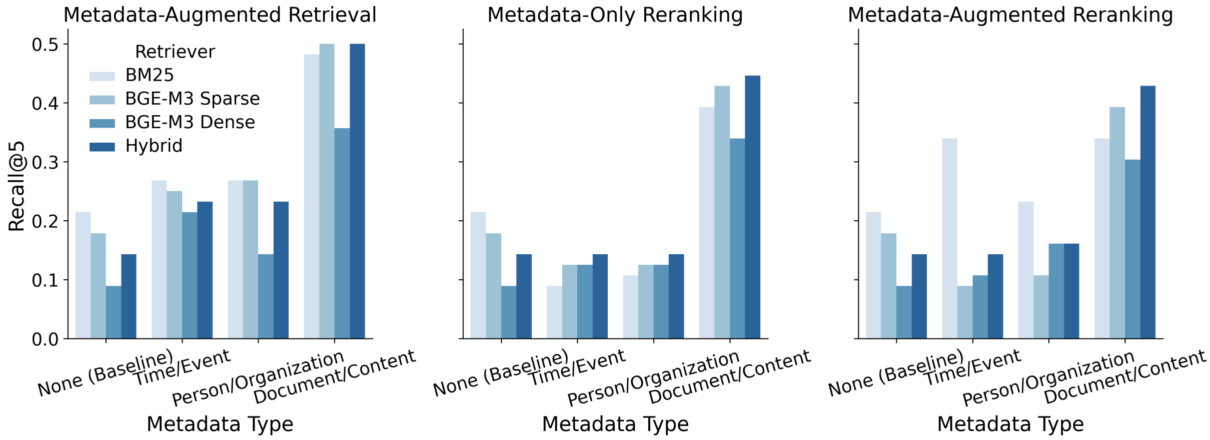


Figure 5: TPCG retrieval performance across different metadata integration stages and metadata types. Left: Metadata-Augmented Retrieval performance across different metadata types. Center: Performance of Metadata-Only Reranking across different metadata types. Right: Retrieval performance of Metadata-Augmented Reranking across different metadata types.

entity queries, the system achieves higher overall evaluation scores on Multi-entity questions. This indicates that the LLM can generate high-quality answers for Multi-entity queries even with partial or imperfect context. In contrast, despite adequate retrieval for Time-focused questions, generation performance remains poor, highlighting that time-sensitive reasoning constitutes a key limitation of the LLM rather than retrieval.

2) Retrieval with Document-level Metadata We examine the role of document-level metadata in the retrieval process, focusing on metadata type and integration stage.

Different Metadata Type. Figure 5 compares TPCG retrieval performance across different retrievers and metadata types: Time/Event, Person/Organization, and Document/Content, at each integration stage, arguing how metadata affects RAG performance. Document/Content metadata provides the largest improvement over the baseline across all strategies, achieving recall scores roughly twice those of the other types, with the highest around 0.5 under the Metadata-Augmented Retrieval setting. This enhanced retrieval supplies essential context to the LLM, improving answer quality and boosting groundedness and relevance, as shown in Table 4. In contrast, Time/Event and Person/Organization metadata exhibit variable effectiveness across integration stages and are insufficient alone for effective reranking, a trend also reflected in the RAG evaluation scores.

Different Metadata Integration. Figure 5 also illustrates retrieval performance across different metadata integration stages. Metadata-Augmented Retrieval consistently outperforms the baseline across all retrievers and metadata types. For BM25, which is used for TPCG, recall increases from 0.21 to 0.48, indicating that integrating metadata directly into document embeddings during retrieval enables the most effective use of structured information.

In contrast, Metadata-Only Reranking produces only modest gains and sometimes underperforms the baseline; for BM25, recall drops from 0.21 to 0.08, suggesting that metadata applied solely at the reranking stage is insufficient. Metadata-Augmented Reranking yields mixed results: while recall generally improves over the baseline, gains are smaller than those of Metadata-Augmented Retrieval, leading to greater instability in generation.

7 Conclusion

This study investigates the application of RAG to historical open-ended question answering using two Traditional Chinese historical datasets, Fort Zeelandia and TPCG, along with query sets. By examining the impact of query types and metadata integration strategies on retrieval and end-to-end RAG, we show that early-stage metadata integration substantially enhances performance. Our results also reveal persistent challenges: hallucinations are frequent during generation, and temporal or multi-hop queries are particularly difficult because of the low retrieval recall. These findings inform future humanities-focused RAG research and underscore the need for robust retrieval strategies in historical and Traditional Chinese contexts.

Acknowledgments

We would like to express our sincere thanks to the National Science and Technology Council (NSTC), Taiwan, for funding this research project under Grant No. NSTC 113-2740-H-002-001-MY3, "TAIHUCAIS: TAIwan HUmanities Conversational AI Knowledge Discovery System". The support has been instrumental in enabling the study.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, et al. 2018. Ms marco: A human generated machine reading comprehension dataset.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759. Association for Computing Machinery.

Jiawei He, Boya Zhang, Hossein Rouhizadeh, Yingjian Chen, Rui Yang, Jin Lu, Xudong Chen, Nan Liu, Irene Li, and Douglas Teodoro. 2025. Retrieval-augmented generation in biomedicine: A survey of technologies, datasets, and clinical applications.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024a. Cmmlu: Measuring massive multitask language understanding in chinese.

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2024b. Biomedrag: A retrieval augmented large language model for biomedicine.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2024. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. 2024. Gpt-4o system card.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Gerard Salton and Chris Buckley. 1987. Term weighting approaches in automatic text retrieval. Technical report, Cornell University, USA.

Rahman S. M. Wahidur, Sumin Kim, Haeung Choi, David S. Bhatti, and Heung-No Lee. 2025. Legal query rag. *IEEE Access*, 13:36978–36994.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Te-Lun Yang, Jyi-Shane Liu, Yuen-Hsien Tseng, and Jyh-Shing Roger Jang. 2025. Knowledge retrieval based on generative ai.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, et al. 2024. Crag – comprehensive rag benchmark.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the Symposium on Computer Science and Law on ZZZ*, CSLAW '25, page 169–193. ACM.

A Appendix

A.1 Fort Zeelandia Dataset Example

Figure 6 gives an example from the Fort Zeelandia dataset.

Query Set

Query: 誰在1632年被派去，再度調查中國漳州河至南澳島的海岸地形？
Question complexity: Single-hop
Entity focus: People
Gold passage ID: 熱蘭遮城日誌/I-C/1632-03-04

Document dataset

Passage ID: 熱蘭遮城日誌/I-C/1632-03-04

Passage content: 2月28，29日，3月1，2，3，4日。無事，為快艇Catwijck號卸貨。
今天有一艘中國人的戎克船從漳州河來到此地，帶消息來說，有兩三艘戎克船裝著貨停泊在廈門港口，準備沒有軍門的通行證也要出航前往馬尼拉，被軍門下令逮捕，貨物全被沒收充公，因此本季前往馬尼拉的船，將不會有二十艘，最多也將不過十四到十五艘小戎克船。
今天也決定了關於長官普特曼斯閣下要搭快艇Catwijck號前往巴達維亞的事，議會重新討論以後，仍然決議，一切準備妥當以後，該快艇就要立刻出航；並決議，為要在本季還能儘快向總督閣下報告日本方面傳來的消息和公司在這邊的消息，決定要在本月底以前備好一艘戎克船，以便到時可立即出航前往巴達維亞；又決議，在5月1日到10日之間，下席商務員Jacob van Sandt要率領兩艘裝備齊全的戎克船前往中國，去重新探查從漳州河到南澳島之間的中國沿岸，然後去該南澳島以南巡航，或停泊在南澳島下面的岸邊等候巴達維亞來的船隻，以便長官閣下回來的時候，得以向他報告所有他離開以後所發生的公司的事情。

Figure 6: A data sample of the Query Set and its relevant passage in the document dataset from the Fort Zeelandia dataset.

A.2 TPCG Dataset Example

Figure 7 gives an example from the TPCG dataset.

Query Set

Query: 臺灣省議會在1990年3月26日的第九期中，如何評估本省與北、高二市在教育品質上的差異？
Gold passage ID: 003-09-01OA-67-6-6-01-01120

Document dataset

Passage ID: 003-09-01OA-67-6-6-01-01120

Passage content: 臺灣省議會公報 第六十七卷 第九期 九一二分校獨立或設校之權限，自去年開始已授權縣市政府。 蘇洪議員月嬌質詢：如果縣府不做呢？ 教育廳陳廳長倬民答復：這是不太可能，因縣長是民選的，對於民衆的需要他一定會重 視。蘇洪議員月嬌質詢： 他人的想法你怎可知道，請教就貴廳的立場是如何？ 教育廳陳廳長倬民答復：只要本廳可爲助力的一定儘量支援。至於縣府不做時，我想本廳 應也是有辦法制肘的，或如將補助經費減少等方式。 蘇洪議員月嬌質詢：■召集開會，請也邀請臺影公司到現場錄影 >以上謝謝。 余議員玲雅質詢：(79年8月14日) 首先請教廳長，我想廳長應知道，本省與北、高二市所受教育的 質差很多。據本席持有之分析資料表，如以福利、設備、退休而言， 本省是遠遠落在北、高二市之後，不知廳長是否了解，甚至其都有不 同標準。尤其高雄市縣只有一線之隔，因此大部分只要有機會轉到高 雄市服務。我們常說當兵、納稅、受教育是人的三大義務，而納稅方 面而言，省民並不亞於北、高二市民 >但爲受教育的處遇就要差很多 ?就此點不知廳長的感想如何？ 教育廳陳廳長倬民答復：省市的教師待遇有差別是事實，但就全省二十一縣市而言，我們也不能保證各縣市教師的福利措施都完全一樣，因這是涉及到各縣市 財政問題。但要以投入教育經費每位學生平均數額而言，本省是比北市低 >不過要比高市略高。至於所提福利措施方面，據資料有一部分 本省是較少，如北高二市有交通費補助、水電費部分補助，這部分在 本省一般地區就沒有，但偏遠地區仍有補助的，所以就說是補助重點 不一樣。另所提因老師的福利差遇，致造成他們的流向都市，我認爲這就 很不公平，就本省而言經常有缺額現象，而北、高二市也有同樣情形 ,但其遇有缺額一經招考即馬上可補足，而這些人員無非是來自本省 ,因而造成我們的教師缺額更形嚴重。以上年度而言本省的教師缺額 僅國小即高達三、00三位，本以師院畢業生及國小師資進修班的一 千七百人，計三千多人應已可補足，但目前還是有缺額，而所缺的應 僅止是退休人員的數目，但北、高二市也缺額，馬上的找去我們的老 師，因而又造成了我們的缺額，這就是造成教育上目前很不平之處。 以上這種現象，我是覺得教育行政單位是有責任，如何的在培育師資 方面做到供過於求，如此的師資才能安定，減少流動性，這也是我們 努力的方向。就目前估計，國小師資到明年大概已可平衡，因此以後流動就會減少。但這個政策有個缺點，即被分發到偏遠地區的自認已 無法調動，屆時恐又會有意見。所以任何政策如何的能取到一個平衡 點，這也是我們努力的方向。至於北高二市福利待遇較好的問題，本廳也多次向教育部反映， 亦受到重視*因此，對本省偏遠地區的老師加給也特別編列一部分經 費，同時其補助辦法、方式本廳也擬定送中央教育部轉報行政院，只 要奉核定，則對本省偏遠地區老師的加給會有很大的改善。余議員玲雅質詢：我也是知道廳長很認眞的在做，但以高雄縣市而言每年的自強活 動經費高市是一千元、高縣是八百元«服裝費高市是二千元，高雄縣...

Start date: 1990-03-26
End date: 1990-03-26
Vloume: 67：第九期，（1990年）
Published date:
Members: 余玲雅
Agencies: 教育處，教育廳
Decree:
Officals:
President:
Title: 臺灣省議會第九屆第一次定期大會：質詢-教育-教育
Abstract: 本省與北、高二市所受教育的質差很多，不知感想如何？延長十二年國教 對學生數的如何分配與現行教育體系下是否足夠分配，均應慎重考慮；目前課程的僵化是值得檢討，尤其課程的一元化是適應於升學的，對於不升學的根本無法接受；如教材要有所改進時，建議應要如何落實本土化教育；如果調整教材之後老師是否要進修？對於新教材要有新的教法？
Type: 公報
Category: 質詢
Subject:
Keywords:
Attachments:
References:
Remarks: 於第九屆第三次臨時大會舉行

Figure 7: A data sample of the Query Set and its relevant passage in the document dataset from the TPCG dataset. Note that some metadata fields are missing in the raw data source, such as **Decree** and **Officials**. The second half of **Passage content** is omitted for brevity.

A.3 Generation Prompt

The full prompt provided to GPT-4o for response generation, given the query, retrieved passages, and available metadata, is shown in Figure 8.

你是一個根據檢索文檔回答問題的語言模型。你會收到一個使用者的問題，以及一筆或多筆可能相關的文檔，每筆文檔都包含「內容（content）」和「後設資料（metadata）」。請遵循以下指引回答問題：
1. 根據文檔內容和後設資料回答問題，避免加入檢索資料中未提及的資訊。
2. 優先使用內容一致、來源可靠的文件，不使用有矛盾、錯誤或無關的資料。
3. 如果所有文件都與問題無關，請誠實地說明找不到答案。

使用者問題
{query}

檢索文檔
內容：{content$_1$}
{metadata$_{1\text{-}1}$ name}：{metadata$_{1\text{-}1}$ content}
{metadata$_{1\text{-}2}$ name}：{metadata$_{1\text{-}2}$ content}
...
{metadata$_{1\text{-}m}$ name}：{metadata$_{1\text{-}m}$ content}

內容：{content$_2$}
{metadata$_{2\text{-}1}$ name}：{metadata$_{2\text{-}1}$ content}
{metadata$_{2\text{-}2}$ name}：{metadata$_{2\text{-}2}$ content}
...
{metadata$_{2\text{-}m}$ name}：{metadata$_{2\text{-}m}$ content}

...

內容：{content$_n$}
{metadata$_{n\text{-}1}$ name}：{metadata$_{n\text{-}1}$ content}
{metadata$_{n\text{-}2}$ name}：{metadata$_{n\text{-}2}$ content}
...
{metadata$_{n\text{-}m}$ name}：{metadata$_{n\text{-}m}$ content}

Figure 8: RAG generation prompt to GPT-4o. Retrieved passages are numbered from 1 to n, representing the 1st retrieved passage to the nth retrieved passage. Metadata rows for each retrieved passage are numbered from 1 to m, representing the 1st type of metadata to the mth type of metadata.

A.4 Evaluation Prompt

The full prompt provided to Gemini-2.5-Pro for response evaluation, given the query, golden passages, retrieved passages, available metadata, and model response of GPT-4o, is shown in Figure 9.

````
你是一位專業的回答品質評估員，負責根據模型回覆是否正確地整合了提
供的【標準答案文檔】與【檢索文檔】來回答問題，進行評分。

請根據以下三個面向，以 0 分為每個面向的初始分數給出各個面向的分
數，最後計算總分：

---

**面向一：【是否包含標準答案文檔的內容】**

* 若檢索文檔中不包含標準答案文檔，則此面向為 0 分。
* 若有包含，把標準答案文檔的內容以最小單位拆解成多個敘述，只要模
型回覆每包含一個敘述，就加 1 分。


**面向二：【是否整合檢索文檔中其他與標準答案文檔相關的內容】**

* 標準答案文檔除外，若其他檢索文檔內容都與標準答案文檔無關，則此
面向為 0 分。
* 若有與標準答案文檔相關的其他文檔，把所有相關的其他文檔的內容以
最小單位拆解成多個敘述，只要模型回覆每包含一個敘述，就加 1 分。
* 出現多個內容重複的敘述只加一次分。

**面向三：【是否包含幻覺或無關內容】**

* 每一個包含幻覺、參考無關文檔、對回答問題沒有幫助的敘述，各減 1
分。

---

請嚴格依據以上標準評分，並輸出以下格式：

```
面向一：+[分數]
面向一評分理由：
[條列每個加分項]

面向二：+[分數]
面向二評分理由：
[條列每個加分項]

面向三：-[分數]
面向三評分理由：
[條列每個減分項]

總分：[分數總和]
```

以下是你要評估的資料：

---

### 問題
{query}

### 標準答案文檔
{golden passages}

### 檢索文檔
{retrieved passages}

### 模型回覆
{model response}
````

Figure 9: RAG evaluation prompt to Gemini-2.5-Pro. Formats for golden passages and retrieved passages are the same as the retrieved passages in the RAG generation prompt.

A.5 Fort Zeelandia Dataset Evaluation Example

Figure 10 gives a detailed example of the evaluation result on a single-hop question from the Fort Zeelandia dataset.

Focusing on the third scoring dimension of the evaluation result, we can observe that GPT-4o, which is used for model response generation, can still hallucinate, even when the golden passage is retrieved as the first retrieved passage. The hallucination may be attributed to the model's tendency not to include violence-related information from the golden passage, resulting in an incomplete response.

A.6 TPCG Dataset Evaluation Example

Figure 11 and 12 give two detailed examples of the evaluation results on the TPCG dataset.

In the first example, the model response from GPT-4o covers almost all the information in the golden passage, which is also the fifth retrieved document. However, the meeting session (in the **Title** metadata field) of the first retrieved document is wrongly linked to the golden passage and appears in the model response. This example suggests the limitation that hallucination may come from the integration of rich and complex metadata information.

In the second example, the evaluation result of the second scoring dimension shows that GPT-4o can still summarize related information from other retrieved passages even though the golden passage is not given for generation. Nonetheless, the model regards the requirements and questions, which are raised by council members, as implemented specific measures, introducing another type of hallucination due to the imprecise interpretation of retrieved passages.

Query

荷蘭當局決議要求牧師甘迪留斯以「甜蜜的方法」勸阻Taccaran前往日本，這反映了當局在處理此事上的態度為何？

Golden passage

Passage ID: 熱蘭遮城日誌/I-F/1633-07-07
Passage content: 7月7日。長官閣下在議會提出上述昨天的意見，乃決議[1]，由於議會與長官閣下都同樣認為，中國人由於恐懼，（跟以前一樣）會再用漂亮的言談來誤導拖延我們，使我們耗費很多費用，因此，我們要遵照總督閣下與東印督議會的指令，要用最猛烈，又儘量少流血的方法，向中國作戰，去攻擊，奪取他們的船隻，為此目的，要派快艇Bredam號、Wieringen號與平底船Warmondt號及一艘戎克船一起部署在南澳下面，去執行要給他們的指令，於其餘的船隻離開他們以後一兩天，要留在那裡，把所有泊在岸邊或擱在陸地上的船隻通通燒毀，把奪得來的貨物保存下來。達成這任務之後，上述快艇要各往預定地點，即Bredam號要偕同一艘戎克船去好望角，Wier-ing號要偕同一艘戎克船去鐘灣[2]，平底船Warmondt號則要去東山，在那裡逮捕中國人的及葡萄牙人的船隻，而大船Middelburch號、快艇Texel號、Weesp號、Couck-ercke號、Catwijck號、Zeeburch號、Salm號、Kemphaen號及戎克船打狗號則要前往漳州河及廈門，要去那裡同樣攻擊燒毀他們的船隻。

並決議，於上個月14日捕獲那艘從薩摩來到大員的那艘chiamboey船的船長，舵工及還在我們手中的水手，都要從那裡送來此地，以儘量避免跟日本人衝突而造成各種災難。

並決議，要用各種方法阻止麻豆社的酋長Taccaran前往日本，因為他在日本的出現，從各方面看來，會造成荷蘭聯合東印度公司很大的不利。據悉，牧師甘迪留斯跟上述Taccaran有過很好的友誼，要請他用甜蜜的方法留住該Taccaran，詳情請看決議錄。

最後決議，因鑑於在奪取戰利品時，經常發生混亂無序的現象，因此要發出公告，貼在每艘快艇上，禁止任何人不得越軌傷害中國人，更不許殺死中國人，每一個人都要繼續拿著武器，靜靜地留在船裡，詳情及如何處罰違規，請參閱該告示。

Retrieved passages

Passage ID: 熱蘭遮城日誌/I-F/1633-07-07
Passage content: 7月7日。長官閣下在議會提出上述昨天的意見，乃決議[1]，由於議會與長官閣下都同樣認為，中國人由於恐懼，（跟以前一樣）會再用漂亮的言談來誤導拖延我們，使我們耗費很多費用，因此，我們要遵照總督閣下與東印督議會的指令，要用最猛烈，又儘量少流血的方法，向中國作戰，去攻擊，奪取他們的船隻，為此目的，要派快艇Bredam號、Wieringen號與平底船Warmondt號及一艘戎克船一起部署在南澳下面，去執行要給他們的指令，於其餘的船隻離開他們以後一兩天，要留在那裡，把所有泊在岸邊或擱在陸地上的船隻通通燒毀，把奪得來的貨物保存下來。達成這任務之後，上述快艇要各往預定地點，即Bredam號要偕同一艘戎克船去好望角，Wier-ing號要偕同一艘戎克船去鐘灣[2]，平底船Warmondt號則要去東山，在那裡逮捕中國人的及葡萄牙人的船隻，而大船Middelburch號、快艇Texel號、Weesp號、Couck-ercke號、Catwijck號、Zeeburch號、Salm號、Kemphaen號及戎克船打狗號則要前往漳州河及廈門，要去那裡同樣攻擊燒毀他們的船隻。

並決議，於上個月14日捕獲那艘從薩摩來到大員的那艘chiamboey船的船長，舵工及還在我們手中的水手，都要從那裡送來此地，以儘量避免跟日本人衝突而造成各種災難。

並決議，要用各種方法阻止麻豆社的酋長Taccaran前往日本，因為他在日本的出現，從各方面看來，會造成荷蘭聯合東印度公司很大的不利。據悉，牧師甘迪留斯跟上述Taccaran有過很好的友誼，要請他用甜蜜的方法留住該Taccaran，詳情請看決議錄。

最後決議，因鑑於在奪取戰利品時，經常發生混亂無序的現象，因此要發出公告，貼在每艘快艇上，禁止任何人不得越軌傷害中國人，更不許殺死中國人，每一個人都要繼續拿著武器，靜靜地留在船裡，詳情及如何處罰違規，請參閱該告示。

Passage ID: 熱蘭遮城日誌/I-G/1635-05-09
Passage content: 5月9日。有一艘戎克船出航前往〔下〕淡水，要去捕魚。有南風，好天氣。

今晚長官閣下接到牧師們從新港寄來的一封信，從那封信的內容得悉，麻豆的一個人Taccaran，他以前有一段長時間在各方面的看法都被我們當作是朋友，而被我們善予款待，最近對我方裝腔作勢，表現非常莽撞，大膽而且高傲，屢對新港及其附近村落的荷蘭人暴怒，說，荷蘭人怕他，因為他們的人殺過他們〔荷蘭人〕的士兵，所以如果要使附和我方的新港人害怕，他們也必須要這樣做；因此新港人非常恐慌起來，都想要對麻豆人作戰，而且聽說，麻豆人要來放火燒毀他們的村子，因為在那裡只不過有10到12個荷蘭人。他們〔牧師們〕也寫說，上述Taccaran很隨意地把一種他們稱之為pockon[1]的器具〔或樂器，instrument〕要拿去Topangh[2]，用以強調，將來他要保護他們。

因為這些事情如果不予及時處理，必將造成公司的侮辱和傷害，因此那些牧師們非常懇切地請求，長官閣下要親自帶領一隊約80到100個士兵〔前來新港〕，因為他們確信，這將使新港人鼓舞起來，而使麻豆人的傲氣消沈下去，並可使其他村落的人保持應有的服從與和平；而且，為要使這些大膽的民族更加害怕起來，長官閣下也要前往目加溜灣（藉口要出去散步），據上述牧師們的見解，這是一趟不必動武器的出征。

Passage ID: 熱蘭遮城日誌/I-G/1635-05-18
Passage content: 5月15，16，17，18日。無特別的事，只有這幾天都很忙著從上述前來的各戎克船收購他們的絲，絲貨及其他貨物，這幾天有幾艘戎克船出航前往中國。

大部份時間吹北風。

今天長官收到牧師羅伯．尤紐斯的一封來信，信裡寫說，新港人表示抱歉，他們不知道他閣下禁止向麻豆人再表現友誼，並請他閣下對此原諒他們；要跟蕭壠人多來往，而跟麻豆人少來往的〔想法〕，他認為那是他閣下很週到的想法，可以把麻豆人壓制下去，而把蕭壠人聯合到我們這邊來，但不要做得使他們看出我們跟他們友好是為要獲得某種利益的。此外又寫說，昨天有兩個麻豆的長老去過那裡，乃向他們陳述他閣下對Taccaran甚為憤怒，並責令他們，任何人都不得妨礙持有公司證件的中國人在魍港燒石灰或捕魚，也不得侮辱他們或其他傷害的舉動。對此他們承諾會遵行，並會向所有的麻豆人傳達。他們因此請求，派兩三個荷蘭人跟他們一起去〔麻豆〕，用以象徵友誼。對此上述尤紐斯予以拒絕說，要等候他閣下命令才能派遣[1]。

Passage ID: 熱蘭遮城日誌/II-F/1644-09-06
Passage content: 9月6日。好天氣，吹陸風。隊長Pieter Boon率領快艇Leeuwercq號與那艘大的小艇出航前往淡水。繼續忙著裝糖桶到Dolphijn號上，要運去交給大船Haerlem號。

我們也要探訪傳道Gerrit Jansz. Hartgringh寫一封信，用以答覆昨天收到的他的來信，要他把學校老師Caesar van Winschooten儘早送來此地，以便來為他所犯的幾樣過錯答辯。

今天有1艘小戎克船出航前往淡水載硫磺，搭31個人；也有1艘coya船出航前往澎湖，空船，搭11個人；另有2艘coya船出航前往中國，載有鹹魚，搭5個人[1]。

今天長官與議會允許並規定，首先，教會議會得以檢查所有教會人員的工作情況，並得予以停職或降職；並規定，該教會議會必須將那些人的犯錯資料以及他們應受處罰的意見交上來，以便於需要時〔福爾摩沙〕議會可針對有關案件進一步審議[2]。fol.171其次，在探訪傳道權力下的所有本地學生，以及在福爾摩沙的學校擔任學校教師的居民〔指原住民〕，都要維持現狀，這些教師或學生，政務員都不得調派去做任何其他工作，除非有緊急的需要[3]。

第三，那些長老，即當地的酋長，將取消允許他們將學生帶離學校的權力（因為還有很多人是異教徒，他們妨礙神的教會的發展），唯一例外的就是新港的長老，在有進一步的決議以前，他們仍得擁有這項權力，因為他們是最虔誠信神的當地人。

也決議，從所提選舉長老與執事的兩倍人選中，決定選擇下席商務員Eduard aux Brebis[4]為長老，下席商務員Wijnant Rutgers[5]為執事。

並決議，要派2個探訪傳道和6個士兵隨牧師Simon van Breen去北區，以便去那邊學習語言，並推展教會的工作。並決議，要派傳道Hans Oloff與Hendrick Veer，取代探訪傳道Gerrit Jansz. Hartgringh，去大木連工作；並派曾經在蔴豆與阿猴（Acau）任職的學校教師Caesar van Winschooten[6]，fol.171v以及曾在目加溜灣任職的探訪傳道Joost Gilles，去新港工作[7]。

並決議，在那艘Dolphijn號裡，除了裝運要帶去交給大船Haerlem號的糖以外，還要裝上要運回祖國的7箱各種布料、4桶薑糖、2桶茯苓、170包瓷器等、100個圓形的大醃缸。

也決議，要出售大量的tacabossen[8]給中國人，用以填補新開始的〔資金的〕短缺。

本月2日偕同稽查官Adriaen van der Burgh從澎湖抵達此地，並來這議會提過他們的請求的那幾個暹邏人，來回答說，他們已經考慮過我們那時回答他們的意見，即自己〔租〕用一艘戎克船去日本，因為我們的船都已經出航了，他們回答說，因為考慮到他們沒有士兵，不能使用中國人的船，現在〔要航往日本的〕季節又快要過去了，還有其他種種阻礙，因此最後認為，應當為了他們的主人，暹邏的國王，再次來向長官請求，讓那一封國王的書信，以及25到30個他們的人員和那些日本的翻譯員，得以用一艘我們的船送往日本。這個請求，長官與議會聽了，並加考慮fol.172認為這樣做會有困難，因此決定，要拒絕他們的請求，禮貌地回答他們說，除了今年〔要航往日本的季節〕已經很晚之外，我們的船也不得運外國人去日本，而且我們也不知道現在日本的和暹邏的國王互相之間關係如何，因為已經數年，暹邏的使臣似乎未被日本的皇帝陛下接見過，他們的人員，禮物和運去的貨物都被拒收，而且那些日本的翻譯員，無可懷疑地，都一定會被處死，而把他們載去日本的我們，也會因而遭遇很大的危險，我們自己也會被砍頭。我們把這些話告訴他們以後，他們感謝議會給他們的警告，說，回去暹邏以後，會這樣向他們的國王報告。於是請求，因為（如上所說）不懂如何搭用中國人船，也沒有信心將國王的書信放在中國人的船裡，因此請求我們，讓他們搭我們的船回去暹邏他們的國王那裡，並請求說，為了該書信的緣故而攜帶的他們的蘇木和其他貨物，准予在此地出售，並幫助他們出售這些貨物。對此，長官與議會決議，要簡要地回答他們說，對於第一個要求，fol.172v即要搭我們的船回去暹邏的事情，我們將予以考慮，對於第二個要求，即要出售他們的貨物的事情，將允許他們，並在不損害公司的利益下，準備要幫助他們，詳情載於今天的決議錄裡[9]。

Passage ID: 熱蘭遮城日誌/III-A/1648-11-02～1650-03-10/補充資料
Passage content: 〔補充資料〕

〔1648年11月2日至1650年3月10日〕

1648年12月裡有2艘船從大員來到巴達維亞入港。平底船Juffer號於12月5日抵達，所載貨物有1,421箱砂糖；平底船Os號於12月21日抵達，也載來一批砂糖和其他數種商品，總值28,681.18.15荷盾。

Dagregister Batavia〔荷文本《巴達維亞城日誌》〕1647-1648，171，188。

雖然中國大陸的內戰還沒結束，中國商人從南方的商港漳州、安海、廈門、金門跟福爾摩沙的交易還相當暢通。謠傳還會有很多暴亂發生。滿州軍隊已經攻取內陸三個城市，殺掠百姓。國家遭遇飢荒，因為農地大都荒蕪。1648年又見無數的中國人逃來福爾摩沙，其中有500個婦女和1,000個小孩。這年福爾摩沙群島上有超過20,000的成年中國人。Generale Missiven〔《總督一般報告》〕1649年1月18日函，1639-1655，Van der Lijn、Caron、Reniers.、Van Dutecom、Demmer，VIII，Batavia，18 januari 1649，354，355。

荷蘭聯合東印度公司在出島和福爾摩沙的商館都賺到很多錢。這些盈餘的錢用銀送往東京和暹邏的商館去收購商品；總計運150箱去東京，20箱去暹邏。

福爾摩沙的長官與議會決議，要稍微調高糖價，用以鼓勵中國農夫種植甘蔗。過去一段時間，那些農夫到處種稻與其他糧食作物，因為中國大陸缺糧，造成糧食作物價格飛揚。公司也看到運糖去日本市場的特別好的機會。長久以來，日本的糖的市場由中國供應，現在他們的供應受到阻礙了，公司正可利用這機會出口福爾摩沙的糖去日本。熱蘭遮城堡的人員很缺乏數種需用品，例如柏油、繩索、錨等物。有兩百多個該島的駐軍契約即將屆滿，因此議長Pieter Anthonisz Overtwater請求要及時派兵來替補。

淡水的議會報告說，北方的噶瑪蘭（Cabalan）人對荷蘭人發怒[1]。直接的起因是隊長Thomas Pedel下令去處罰一個罪犯所引起的。此外，從北方地區傳來的都是好消息。那十二個龜崙（Coeland）的村社，已有十一個村社的長老跟公司結盟了。位於本島東岸的哆囉滿（Tarraboan）地區也只剩下一個反抗的村社。近期中，將派下席商務員Anthony Plockhoy去該村社勸和，並將沿途繪製該山區的地圖。

1648年春天曾有兩個福爾摩沙的村社首長從北方去熱蘭遮城堡探望。雖然他們被接待時，特別稱讚過公司的政策，但以後卻批評對他們的孩童的基督教教育。跟在城堡附近的學校的授課情形比較起來，他們那地區的教育品質就差的遠了。牧師Jacobiis Vertrecht去虎尾壠附近幾個村社探訪，說服了那裡兩個地方的頭人跟公司締和。他相信，那兩個村社的人到時會派代表來參加下次的地方會議。

〈議長P. A. Overtwater致總督C. van der Lijn與巴達維亞議會函〉，大員，1649年2月1日。VOC 1172，443-449[2]。

長官Nicolaas Verburch於1649年11月18日寫說，1649年6月18日到8月19日從巴達維亞派出，經由暹邏、占碑、東京和日本航來目的地大員的十八艘平底船，有十四艘抵達大員了。還沒看到平底船Campen號、Witte Paert號、Gulden Gans號與Salm號。希望這幾艘船已經找到避風港安全渡過這颱風季節。上述公司的奴隸那些Pampang人〔指逃走被捕捉的那些人〕，由總督府裁定釋放，因為知道福爾摩沙非常缺乏勞工。因為中國一向交易的貨物，例如絲和瓷器，缺貨，所以現在用黃金來交易。長官Verburch認為，這種商品的短缺是中國內部繼續動亂的徵候。很多來福爾摩沙的中國商人也同意這種看法。這種不穩定的中國貿易，使大員的公司當局更難於供應祖國、巴達維亞與印度沿岸通常的需求。Verburch是想要聽從巴達維亞的命令，即要將中國的黃金降價到每十兩24卡拉的黃金兌一百兩銀。不過他看不出有將商品價格按照比例降價的可能性，因為他不敢向福爾摩沙的中國商人提出降低黃金價格的事情。公司已經向他們說過，目前黃金價格不會低於每十兩24卡拉的黃金兌一百一十五兩精銀。如果這諾言不履行，黃金將可能不再運來福爾摩沙了。不過，所提每十兩24卡拉的黃金兌一百兩精銀的價格，在隔年的大員帳簿就有記載了。這樣獲得的利益可能可以補償在Coromandel和公司其他商館的損失。

想要鼓勵農夫在福爾摩沙生產絲的嘗試，沒有獲得預期的效果。長官提出相對的辦法，就是把中國絲的進價提高到每擔七百里爾。福爾摩沙的農地，因為現在不必再去種植桑樹，可以補償那提高的價格。中國贌農〔pachtboeren可能是指向公司付租金種田的農夫，如同清代的墾首〕抱怨說，公司的士兵在收稅和分發人頭稅單時態度粗暴。議會答應將注意改善，以防類似事情發生。按照長官Verburch的看法，公司與這島上中國移民之間和諧的關係是非常重要的；他認為中國人是：福爾摩沙島上唯一提供蜂蜜的蜜蜂，沒有這些人，尊貴的公司是無法在此生存的[3]。

中國人湧入本島的潮流已見減緩。1650年11月登記居留的中國人為11,339人，其中有838個女人。福爾摩沙的作物，過去這季節因乾旱而歉收。同時，〔公司人員〕在出島私自交易的事情曝光，遠超過了合理的界限。以後在大員與暹邏的轉運站裝船時，對目的地為東京與日本的船將嚴格檢查。

〈長官N. Verburch致總督C. van der Lijn函〉，大員，1649年11月18日。VOC 1172，466-491。

大員商館在1648-1649年會計年度淨賺了467,000里爾。福爾摩沙的長官報告說，養蠶業大爲倒退。他建議總督府當局對養蠶業的前景要趕快做決定性的決策。在巴達維亞的人認爲，在福爾摩沙島養蠶並非不適，而是被中國人的頭領暗中破壞了，他們害怕公司對這行業變成太內行。

現在中國嚴重的飢荒已經過去，逃來福爾摩沙的中國人，有些人又回去他們在中國的老家了。1649年春天繳納人頭稅的中國人爲一萬兩千人，而1648年3月還有一萬四千多人繳納。駐軍有984個軍人。由於稻米和其他糧食作物在中國市場降價了，所以本島的農夫越來越多人種植甘蔗。公司預計，可自鄰近村社的農地收到作物什一稅46,000里爾。長官Verburch指出，原住民屢受中國人贌商壓榨，令人憂慮。那些中國人利用他們進口舶來品的專利，完全隨意喊價，福爾摩沙當地人爲此付出過分多餘的費用，而他們提供的鹿肉和鹿皮，則必須無奈地按照通常價格交易。Verburch警告說，如果公司對這失常狀況不迅速加以處理，必將造成原住民的騷動。

今年又有很多派駐南區村社的公司人員感染地方病。不過議會還是決定不撤出該地區。佈道活動頗有進展。牧師Simon van Breen於返回祖國之際，隨身攜帶了一些有荷蘭文以及福爾摩沙語文翻譯的宗教經文的稿件。他的任務是，要在荷蘭共和國印刷這些稿件，以便將來在該島可用來作爲基督教教育的教材。

巴達維亞總督府不贊成十七人董事會所提派商務代表去北京訪問滿州朝廷的意見。巴達維亞當局認爲，這個新的中國的當權者不會從永久的政權基礎建立起一有力的中央政府。那些滿州人大事搶掠之後，會再退回去他們在北方的祖國，他們如此期待著。不想派正式的使臣去中國的另一理由是，南方幾個商港都還在忠於明朝的人的手裡。鄭成功是公司與滿州朝廷之間建立任何來往關係的一個直接的威脅。總督與議會判斷，公司與中國之間的商務關係，最好在正式的外交關係之外實際進行。類此政策對商務最無損害，也不致影響中國人對公司運來的日本銀的需求。

《Generale Missiven》〔《總督一般報告》〕1639 -1655，Van der Lijn、Caron、Reniers、Demmer、Barendsz，X，巴達維亞，1649年12月31日，391-395。

Model response

根據檢索到的文檔，荷蘭當局決議要求牧師甘迪留斯以「甜蜜的方法」勸阻麻豆社的首長Taccaran前往日本，這反映了當局在處理此事上的態度是希望以和平和友好的方式來解決潛在的外交問題。當局認為Taccaran在日本的出現可能會對荷蘭聯合東印度公司造成不利，因此希望通過已有的友誼和柔和的方式來達成目的，而不是採取強硬或對抗的手段。這顯示出當局在處理此類國際和外交事務時，傾向於使用非暴力和協商的方式。

Evaluation result

面向一：+4
面向一評分理由：
[+1] 模型回覆提及，要勸阻麻豆社的酋長Taccaran前往日本。
[+1] 模型回覆提及，Taccaran在日本的出現可能會對荷蘭聯合東印度公司造成不利。
[+1] 模型回覆提及，要利用牧師甘迪留斯與Taccaran已有的友誼。
[+1] 模型回覆提及，要使用「甜蜜的方法」（柔和的方式）來達成目的。

面向二：+0
面向二評分理由：
[0] 模型回覆未整合其他檢索文檔中的相關內容。例如，在同一份決議中，荷蘭當局才剛決議要用「最猛烈」的手段攻擊中國船隻，這與對待Taccaran的「甜蜜的方法」形成鮮明對比，能更完整地反映當局務實、依對象而調整手段的態度，但模型並未提及此點。

面向三：-1
面向三評分理由：
[-1] 模型回覆的結論「這顯示出當局在處理此類國際和外交事務時，傾向於使用非暴力和協商的方式」為不恰當的推論。根據同一份【檢索文檔】，當局同時也決議要用「最猛烈」的方式攻擊中國船隻，可見其並非普遍「傾向於使用非暴力和協商」，而是根據對象與情勢採取不同手段。此結論忽略了文檔中的其他資訊，對回答問題沒有幫助。

總分：3

Figure 10: Evaluation result on the Fort Zeelandia dataset.

Query

臺灣省議會在第九期中，對現行教育課程的一元化問題提出了哪些檢討與改進建議？

Golden passage

Passage ID: 003-09-01OA-67-6-6-01-01120
Passage content: 臺灣省議會公報 第六十七卷 第九期 九一二
分校獨立或設校之權限，自去年開始已授權縣市政府。 蘇洪議員月嬌質詢：
如果縣府不做呢？ 教育廳陳廳長倬民答復：
這是不太可能，因縣長是民選的，對於民衆的需要他一定會重 視。
蘇洪議員月嬌質詢： 他人的想法你怎可知道，請教就貴廳的立場是如何？ 教育廳陳廳長倬民答復：
只要本廳可爲助力的一定儘量支援。至於縣府不做時，我想本廳 應也是有辦法制肘的，或如將補助經...
Start date: 1990/03/26
End date: 1990/03/26
Volume: 67：第九期，（1990年）
Members: 余玲雅
Agencies: 教育處，教育廳
Title: 臺灣省議會第九屆第一次定期大會：質詢-教育-教育
Abstract: 本省與北、高二市所受教育的質差很多，不知感想如何？延長十二年國教　對學生數的如何分配與現行教育體系下是否足夠分配，均應慎重考慮；目前課程的僵化是值得檢討，尤其課程的一元化是適應於升學的，對於不升學的根本無法接受；如教材要有所改進時，建議應要如何落實本土化教育；如果調整教材之後老師是否要進修？對於新教材要有新的教法？
Type: 公報
Category: 質詢

Retrieved passages

Passage ID: 003-09-09OA-75-6-6-01-01058
Passage content: 臺灣省議會公報 第七十五卷 第十一期
灣省各級學校員生消費合作社改進要點」第二條規定：各級學 校均應設置員生社，其有特殊情形，報經主管機關核准者得不 設立。
㈡爲因應實際需要，各校代訂學生餐盒自八十二學年度第二學期 起由學校員生社代辦並得酌收處理費，惟至多不超過餐盒進價 百分之五以內爲限，並以進銷貨登帳方式辦理。至於販售之物 品，應經社務會議決定並經校長同意後，方可出售。
㈢爲加強督導員生社業務...
Start date: 1994/01/17
End date: 1994/12/16
Volume: 75：11，（1994年）
Members: 楊文欣
Agencies: 教育廳
Title: 臺灣省議會第九屆第九次定期大會：質詢-教育-教育
Abstract: 一、全省危險教室知多少？（一）七月十日提姆颱風造成教室倒塌或危險程度？教育廳有無立即調查？有無補救措施？截至目前好像全無資訊。（二）多年前全省危險教室調查有三千間之多，也曾由中央及省府編列預算執行改善，但最近發表全省危險教室還是三千多間，究竟這些年間省教育廳對危險教室做了些什麼？錢用了多少？為什麼沒有效果請陳廳長詳細說明。二、最近的全國教育會議有什麼結果？（一）對教育方針有何改進意見？（二）對中、小學教材有何改進意見？（三）省教育廳對會議有什麼意見提出？三、國中國小教師爭相申請退休，以爭取在退休待遇修改之前退休獲得較好的權益，教育廳對此現象有無因應之道？如何對付？四、國中國小教師的調動以前都是鄉村往都市跑，現在反過來都市的紛紛向鄉村跑，但申請的多，如願的太少，這現象是否會影響教學品質？廳長對此有無對策，儘量人地相宜。五、國中國小房地被占為軍用情形有無改善？請廳長說明。

Type: 公報
Category: 質詢

Passage ID: 003-04-07OA-25-1-6-01-00388
Passage content: 報公會議浦灣凄
期十第 卷五十二第 刊週
特
目
\教育廳潘廳長振球工作報街
新聞處周處長天固工作報告
錄
上一一一一一五
三三九
交通處陳處長來甲工作報告.........S一
一會議紀錄
-第四屆第七次大會第二十三次會議紀錄.三空八
-第四屆第七次大會第二十四次會議紀錄.S七
-第四屆第七次大會第二十五次會議紀綠.一一茜八
質詢及答覆
一書面質詢及答覆.................三兕
民...
Start date: 1961/06/30
End date: 1961/06/30
Volume: 25：第十期，（1961年）
Agencies: 教育處，教育廳
Title: 臺灣省議會第四屆第七次定期大會第二十次會議：報告-教育-教育處、教育廳
Abstract: 教育廳潘廳長振球工作報告：情勢的分析：第一、政府為應事實需要，適時推行九年國民教育；第二、社會型態正由農業社會轉變為工商業社會；第三、國家建設的快速進展，工商企業界及社會各部門，對人力的需求，素質既須不斷提高；新設省立教育學院具有多重教育目標及特色：其一為學制富有彈性、其二為學生以自費為原則，政府毋需負擔鉅額公費；新關係與新內容：中等學校的調整，包括隸屬系統的調整及高中高職發展比例的調整二者，至於高中、高職發展比例的調整，係由中央衡量國家建設需要，作政策性的決定；在設備的充實方面：省府為充實職校設備，除逐年寬列修建設備費外，並申請中美基金補助；結語：第一、時代變動不居，社會進展快速；第二、國際關係日趨密切，人類交往日益頻繁；第三、國家民族正面臨非常關頭，本省復為主要復興基地。
Type: 公報
Category: 報告

Passage ID: 003-06-06OA-44-1-2-07-01934
Passage content: 臺灣省議會公報 第四十四卷
特

一、人事處蔡處長經工作報告
、中華民國六十九年九月二日ノ
厂第六屆第六次大會第一次會議J 議長、副議長*务位議員女士、先生： 今天欣逢貴富六屆第六次大會，經依照大會議程, 列席報告本處半年來辦 理人事書概況，深感榮幸！ 人事ff.政主要任務，在配合省政建設整體發展，支援各機關業務需要，並以 最經濟有效人力，作高度的發揮運用，達到健全組織 > 提高行政功能的目的。各 ...
Start date: 1980/09/02
End date: 1980/09/02
Volume: 44：09，（1980年）
Members: 蔡長經
Agencies: 人事處
Title: 臺灣省議會第六屆第六次定期大會第一次會議：報告-民政-人事處
Abstract: 一、人事處處長蔡長經工作報告
Type: 公報
Category: 報告

Passage ID: 003-04-07OA-25-6-8-00-02275
Passage content: 臺灣省議會公報 第二十五卷 第二十五期
來守交通規則，自己生命自己保，不知陳主席的高見如何？
六、　　鬧得風風雨雨的高雄港務局官員，套購新生地圖利的舞弊巨案，竟雷大雨小 ,不了了之，玆有幾點疑問請教於下：
1據新聞報導，套購案的主要涉嫌官員劉宇强(地畝課長)林高煌(課員)等 經調查站查出違法事實證據，移送高雄地檢處，檢察官偵查屬實起訴後，劉 宇强等不願替人受過，檢具李局長批准出售新生地的公文照...
Start date: 1971/05/24
End date: 1971/05/24
Volume: 25：第二十五期，（1971年）
Members: 涂麗生
Agencies: 臺灣省政府
Title: 臺灣省議會第四屆第七次定期大會：質詢-總質詢-總目
Abstract: 一、省政措施之革新；二、加強社會福利措施；三、目前的教育問題；四、為高雄市八十五萬市民向省主席請命：１請省府協助解決高雄市嚴重水荒；２請在旗津中洲上竹里沿岸建築二噸小型漁船避風港；３請飭令省都市計劃委員會，對於高雄市擴大都市計劃，採納合理的民意重新規劃，並以市價徵收民地；４請高雄硫酸錏公司對生產工人房租津貼比照一般機關學校技工標準發給；５有關高雄市三民區中華橫巷唐榮公司土地處理案；五、希在臺中設立聯絡中心，每個單位設置聯絡員；六、建議本會的建議案，應分類整理加以說明。
Type: 公報
Category: 質詢

Passage ID: 003-09-01OA-67-6-6-01-01120
Passage content: 臺灣省議會公報 第六十七卷 第九期 九一二
分校獨立或設校之權限，自去年開始已授權縣市政府。 蘇洪議員月嬌質詢：
如果縣府不做呢？ 教育廳陳廳長倬民答復：
這是不太可能，因縣長是民選的，對於民衆的需要他一定會重 視。
蘇洪議員月嬌質詢： 他人的想法你怎可知道，請教就貴廳的立場是如何？ 教育廳陳廳長倬民答復：
只要本廳可爲助力的一定儘量支援。至於縣府不做時，我想本廳 應也是有辦法制肘的，或如將補助經...
Start date: 1990/03/26
End date: 1990/03/26
Volume: 67：第九期，（1990年）
Members: 余玲雅
Agencies: 教育處，教育廳
Title: 臺灣省議會第九屆第一次定期大會：質詢-教育-教育
Abstract: 本省與北、高二市所受教育的質差很多，不知感想如何？延長十二年國教　對學生數的如何分配與現行教育體系下是否足夠分配，均應慎重

考慮；目前課程的僵化是值得檢討，尤其課程的一元化是適應於升學的，對於不升學的根本無法接受；如教材要有所改進時，建議應要如何落實本土化教育；如果調整教材之後老師是否要進修？對於新教材要有新的教法？
Type: 公報
Category: 質詢

Model response

在臺灣省議會第九屆第九次定期大會中，針對現行教育課程的一元化問題，提出了以下檢討與改進建議：

1. 課程的一元化主要是為了適應升學需求，但對於不升學的學生來說，這樣的課程設計並不合適。因此，課程的僵化問題值得檢討。
2. 建議教材的改進應落實本土化教育，以更好地適應不同學生的需求。
3. 在調整教材後，教師可能需要進修，以便適應新教材並採用新的教學方法。

這些建議旨在改善現行教育體系，使其更具彈性和多樣性，以滿足不同學生的需求。

Evaluation result

面向一：+4
面向一評分理由：
[+1] 模型回覆提及，要勸阻麻豆社的酋長Taccaran前往日本。
[+1] 模型回覆提及，Taccaran在日本的出現可能會對荷蘭聯合東印度公司造成不利。
[+1] 模型回覆提及，要利用牧師甘迪留斯與Taccaran已有的友誼。
[+1] 模型回覆提及，要使用「甜蜜的方法」（柔和的方式）來達成目的。

面向二：+0
面向二評分理由：
[0] 模型回覆未整合其他檢索文檔中的相關內容。例如，在同一份決議中，荷蘭當局才剛決議要用「最猛烈」的手段攻擊中國船隻，這與對待Taccaran的「甜蜜的方法」形成鮮明對比，能更完整地反映當局務實、依對象而調整手段的態度，但模型並未提及此點。

面向三：-1
面向三評分理由：
[-1] 模型回覆的結論「這顯示出當局在處理此類國際和外交事務時，傾向於使用非暴力和協商的方式」為不恰當的推論。根據同一份【檢索文檔】，當局同時也決議要用「最猛烈」的方式攻擊中國船隻，可見其並非普遍「傾向於使用非暴力和協商」，而是根據對象與情勢採取不同手段。此結論忽略了文檔中的其他資訊，對回答問題沒有幫助。

總分：3

Figure 11: First example of evaluation result on the TPCG dataset. For brevity, part of **Passage content** and empty metadata fields for each passage are omitted.

Query

臺灣省議會在討論精省工作時，省府團隊配合的具體措施有哪些？

Golden passage

Passage ID: 003-10-08OA-84-6-8-00-02658
Passage content: 一、　　省府員工權益自救會將於十月八日北上立法院陳情，過去只有民 眾才會走上街頭抗議, 現在走上街頭陳情的卻是公務人員，對於此 種改變，請問省長有何看法？
二、　　請問省長：省府員工將何去何從？您將如何向中央爭取省府員工 的「工作權」？
省政府87・12・19八七府人一字第一七五九一八號書面答復：
一、　　有關省府員工自救聯盟北上立法院陳情活動，公務人員如果不能 在體制內相關管道反應意見...
Start date: 1998/08/31
End date: 1998/11/06
Volume: 84：16，（1998年）
Members: 徐慶元，陳明文
Agencies: 臺灣省政府
Title: 臺灣省議會第十屆第八次定期大會：質詢-總質詢-總目
Abstract: 有關省長是否秉持「山本五十六」精神所提意見未獲採納，而對於會議決議仍然全力以赴，願意率領省府團隊，配合立法院通過「精省暫行條例」，執行精省工作，使精省的陣痛減至最低？
Type: 公報
Category: 質詢

Retrieved passages

Passage ID: 003-10-07OA-84-6-8-00-01119
Passage content: 北近郊污水下水道系統」，因而接管率暫爲零，近期內即將可提昇 接管率至三・一 %, 八十八及八十九年度將再編列預算繼續辦理，以加速提昇臺北縣污水下水道接管率，並使三重、蘆洲地區先行獲 致提昇居住品質成效。
四、　　「淡水河系統污染整治先期工程」完成後，可暫時達成淡水河不 發臭(無缺氧)之目標，行政院環保署已於八十七年二月報奉行政 院核定繼續推動「淡水河系污染整治計畫後續實施方案」，期冀於 民國九...
Start date: 1998-03-20
End date: 1998-06-26
Volume: 84：10，（1998年）
Members: 劉文雄
Agencies: 臺灣省政府
Title: 臺灣省議會第十屆第七次定期大會：質詢-總質詢-總目
Abstract: 一、「凍省」造成社會不安，政府有無因應對策？二、精省在即，但精省後省政府的地位、各廳處如何調整或整併？既有員工如何輔導轉業或優惠資遣？省政府是否已有完整的配套方案。
Type: 公報
Category: 質詢

Passage ID: 003-10-07OA-84-6-4-01-00318
Passage content: 二、　　大陸進口砂辦理情形。
三、　　多考量原住民保留地開採陸上砂石。
建設廳87・ 5 ・21八七建礦字第四二二一三號書面答復：
一、　　加強疏浚河川、增加河川砂石料源之意見，本廳很贊同, 將建請 水利單位配合辦理。

二、　大陸進口矽石經濟部國貿局於八十六年六月十日公告開放大陸矽 石間接進口。自八十六年七月五日至八十七年四月止共進口二十四 航次，運量約三十一萬公噸，分別在基隆港及...
Start date: 1998-03-20
End date: 1998-06-26
Volume: 84：03，（1998年）
Members: 林宗男，邱茂男，張學舜，王兆釧
Agencies: 建設廳
Title: 臺灣省議會第十屆第七次定期大會：質詢-建設-建設
Abstract: 精省後省府組織，有無將建設廳規劃在內。
Type: 公報
Category: 質詢

Passage ID: 003-10-08OA-84-1-2-04-00719
Passage content: 五、　針對運送危險物品車輛建議應加強相關之修法與安全管理策略研 議。
六、　今後仍應加強易肇事路段之工程改善與嚴格執法，以降低事故之 發生。
七、　爲確實瞭解肇事原因，俾研議事故防制對策，建議員警應加強事 故調查表之資料塡寫，並落實傷者二十四小時之追蹤作業。
陸、結 語
省政交通各項軟硬體建設爲國家發展之根本，不論未來省政府體 制如何變革，照顧與增進省民同胞之交通福祉不可一日終止...
Start date: 1998-08-31
End date: 1998-08-31
Volume: 84：05，（1998年）
Members: 石曜堂
Agencies: 衛生處
Title: 臺灣省議會第十屆第八次定期大會第二次會議：報告-民政-衛生局、衛生處
Abstract: 衛生處處長石曜堂報告
Type: 公報
Category: 報告

Passage ID: 003-10-08OA-84-6-8-00-02661
Passage content: 臺灣省議會公報 第八十四卷 第十六期
作的接續能夠順暢而不致影響省民福祉0
周鄭盧鐘 議 員
錫金秀紹 瑋玲燕和 聯
4
有關精省後省府員工權益保障問題：
一、　雖然精省條例經立法院三讀通過，但其細部規劃作業尚未定案， 建請省府據理力爭，不要放棄爭取員工權益的機會。
二、　在精省過程中，目前尚有省長爲大家爭取權益，但將來省長卸任 後，省府員工若面臨困難，將無處申訴。因此請省長向中央反映...
Start date: 1998/08/31
End date: 1998/11/06
Volume: 84：16，（1998年）
Members: 林進春，張明雄，呂進芳
Agencies: 臺灣省政府
Title: 臺灣省議會第十屆第八次定期大會：質詢-總質詢-總目
Abstract: 有關精省後省府員工權益保障問題：一、日前報載中央分三階段完成精省，不知省府員工權益保障是否會受影響？省議會員工是否也一體適用該退休優惠辦法？二、建請省長向中央薦舉吳副省長及賴副省長為精省後官派省主席人選。
Type: 公報
Category: 質詢

Passage ID: 003-10-07OA-84-6-8-00-01122
Passage content: 業。
2優惠資遣部分
凡經核定辦理資遣者，最高加發十二個月俸給總額之慰 助金(但服務年資仍待商確)·，於業務移轉、移發或業務無 承受機關時，未同時辦理資遣者，每遲延一個月辦理資遣， 減發一個月俸給總額之慰助金，依任職月數遞減至一年期滿 ，不再發給，並強制資遣。
上述決議及相關員工權益保障事項本府業已舉辦說明會轉知所屬 廳處會員工。(最後結論仍待送行政院精省委員會討論通過始確定 。）謝議員言信質詢...
Start date: 1998/03/20
End date: 1998/06/26
Volume: 84：10，（1998年）
Members: 盧逸峰
Agencies: 臺灣省政府
Title: 臺灣省議會第十屆第七次定期大會：質詢-總質詢-總目
Abstract: 請省府化解與中央對精省架構的歧見。省版精省暫行條例和中央規劃相異，造成省府與地方對立甚至抗爭，主要問題在於：（一）省府堅持精省後省仍為公法人。（二）省府主導精省業務。（三）省府保留財產權等，省府對此嚴重歧異部份將以兩案併陳方式在精省委員會進行討論。據瞭解，中央黨部秘書長章孝嚴將前來省府溝通，則宋省長在溝通時化解中央與省之間的歧見。
Type: 公報
Category: 質詢

Model response

根據檢索到的文檔，臺灣省議會在討論精省工作時，省府團隊的具體措施包括：

1. **化解與中央的歧見**：省府被要求化解與中央對精省架構的歧見，特別是在省版精省暫行條例和中央規劃相異的情況下，這造成了省府與地方的對立甚至抗爭。省府堅持精省後省仍為公法人，主導精省業務，並保留財產權等。這些嚴重歧異部分將以兩案併陳方式在精省委員會進行討論。

2. **員工權益保障**：有關精省後省府員工的權益保障問題，省府需要確保員工的權益不受影響，並考慮省議會員工是否適用退休優惠辦法。

這些措施顯示省府在精省過程中需要處理與中央的協調問題以及保障員工的權益。

Evaluation result

面向一：+0
面向一評分理由：
模型回覆完全沒有提及標準答案文檔中的任何內容，例如：
*　率領省府團隊
*　配合立法院通過「精省暫行條例」
*　執行精省工作
*　使精省的陣痛減至最低

```
面向二：+8
面向二評分理由：
模型回覆整合了其他檢索文檔中與問題相關的內容，包含：
*  [+1] 省府被要求化解與中央對精省架構的歧見。
*  [+1] 省版精省暫行條例和中央規劃相異，造成了省府與地方的對立甚至抗爭。
*  [+1] 省府堅持精省後省仍為公法人。
*  [+1] （省府堅持）主導精省業務。
*  [+1] （省府堅持）保留財產權。
*  [+1] 嚴重歧異部分將以兩案併陳方式在精省委員會進行討論。
*  [+1] 有關精省後省府員工的權益保障問題。
*  [+1] 考慮省議會員工是否適用退休優惠辦法。

面向三：-1
面向三評分理由：
*  [-1] 模型將檢索文檔中向省府提出的「質詢」或「要求」（如「請省府化解...」、「有關...問題」）詮釋為省府團隊已執行的「具體措施」，此
為對文檔內容的錯誤解讀，與事實不符。

總分：7
```

Figure 12: Second example of evaluation result on the TPCG dataset. For brevity, part of **Passage content** and empty metadata fields for each passage are omitted.

Method	Metadata Type	Mean Δ	p-value	Significant
Metadata-Augmented Retrieval	Time/Event	0.3571	0.0327	✓
	Person/Organization	0.4464	0.0619	✗
	Document/Content	1.1407	0.0007	✓
Metadata-Only Reranking	Time/Event	-0.3929	0.9175	✗
	Person/Organization	-0.1607	0.7156	✗
	Document/Content	0.8571	0.0005	✓
Metadata-Augmented Reranking	Time/Event	0.5357	0.0036	✓
	Person/Organization	0.2500	0.1095	✗
	Document/Content	0.6429	0.0047	✓

Table 5: Wilcoxon signed-rank test results comparing each retrieval method and metadata type against the baseline for Groundedness on TPCG. The table shows the mean difference (Δ), p-value, and whether the improvement is statistically significant at $p < 0.05$.

A.7 RAG Groundedness Significance Test

Table 5 presents the detailed results of significance testing for the Groundedness metric. For each combination of method and metadata type, we report the mean difference compared to the baseline, the corresponding p-value from the Wilcoxon signed-rank test, and a visual indicator of statistical significance. The results show that the Document/Content metadata type provides the most substantial benefit across retrieval stages, and among the methods, Metadata-Augmented Retrieval with Document/Content metadata achieves the largest mean difference, indicating the strongest improvement over the baseline.

Bridging Underspecified Queries and Multimodal Retrieval: A Two-Stage Query Rewriting Approach

Szu-Ting Liu, Wen-Yu Cho, Hsin-Wei Wang, Berlin Chen
National Taiwan Normal University
{szutingliu, wenyu, hsinweiwang, berlin}@ntnu.edu.tw

Abstract

Retrieval-Augmented Generation (RAG) has proven effective for text-only question answering, yet expanding it to visually rich documents remains a challenge. Existing multimodal benchmarks, often derived from visual question answering (VQA) datasets, or large vision-language model (LVLM)-generated query-image pairs, which often contain underspecified questions that assume direct image access. To mitigate this issue, we propose a two-stage query rewriting framework that first generates OCR-based image descriptions and then reformulates queries into precise, retrieval-friendly forms under explicit constraints. Experiments show consistent improvements across dense, hybrid and multimodal retrieval paradigms, with the most pronounced gains in visual document retrieval—Hits@1 rises from 21.0% to 56.6% with VDocRetriever and further to 79.3% when OCR-based descriptions are incorporated. These results indicate that query rewriting, particularly when combined with multimodal fusion, provides a reliable and scalable solution to bridge underspecified queries and improve retrieval over visually rich documents.

Keywords: RAG, Query Rewriting, Visually Rich Documents, LVLMs, Information Retrieval, Multimodal Retrieval, Optical Character Recognition (OCR)

1 Introduction

Retrieval-Augmented Generation (RAG) has become a central paradigm for building knowledge-intensive QA systems (Gao et al., 2023; Cheng et al., 2025), where large language models (LLM) are paired with retrieval modules to ensure a factual foundation and a broader domain coverage. In text-only settings, such as Wikipedia, news archives, or enterprise databases, RAG systems have been extensively studied, with dense, sparse, and hybrid retrievers achieving strong performance on well-established benchmarks (Lewis et al., 2020; Pan et al., 2022; Abdallah et al., 2025; Sawarkar et al., 2024).

Many real-world enterprise documents are visually rich, containing tables, charts, diagrams, and layout-dependent structures such as those found in product manuals, engineering drawings, or quality control reports. In these cases, relying solely on OCR text extraction leads to partial information loss, as complex visual semantics cannot be fully captured, limiting the effectiveness of traditional text-based RAG pipelines (Appalaraju et al., 2021; Xu et al., 2020). Developing effective RAG systems for such documents requires appropriate datasets and evaluation settings. However, most existing multimodal benchmarks originate from VQA tasks (Tanaka et al., 2025; Wang et al., 2025) are automatically constructed by prompting LVLMs to generate multiple queries for each image, which are then aggregated to form large-scale query-image datasets. These datasets often contain underspecified queries (e.g. 'What does this figure show') that presuppose direct image access; In retrieval settings where only textualized or embedding-based representations are available, such queries fail to identify the correct document reliably, leading to poor retrieval performance.

This limitation motivates our study. We propose a two-stage query rewriting framework that leverages OCR-informed context to reformulate underspecified queries in visually rich RAG settings. Our approach enriches query semantics and produces retrieval-friendly reformulations that better align with multimodal document representations. Extensive experiments demonstrate consistent gains across retrieval paradigms, particularly in multimodal

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 63–70
November 20-22, 2025 ©2025 Association for Computational Linguistics

settings. Our main contributions are as follows:

- We formalize the problem of underspecified queries in multimodal RAG systems.
- We propose a two-stage query rewriting framework that uses OCR-informed image descriptions and prompt constraints to produce retrieval-friendly queries.
- Our evaluations show that query rewriting significantly improves retrieval performance on visually rich documents.

2 Related Work

2.1 Optical Character Recognition

PaddleOCR PP-OCRv5 [1] (Cui et al., 2025) is an open-source multilingual OCR system supporting Simplified Chinese, Traditional Chinese, Chinese Pinyin, English, Japanese, and over 80 additional languages. It follows a three-stage pipeline of text detection, direction classification, and text recognition. Compared with PP-OCRv4, PP-OCRv5 reports a 13-percentage-point improvement in end-to-end benchmark accuracy and includes enhancements for challenging cases such as handwritten text, vertical text, and complex document layouts. Other widely used open-source OCR systems include docTR [2] and EasyOCR [3]. PP-OCRv5's open-source availability and comprehensive documentation make it a practical choice for research and production use.

2.2 Multimodal Document Retrieval

Retrieval-augmented generation (RAG) retrieves external knowledge to enhance large language models (Lewis et al., 2020), but most prior work assumes text-only corpora. Recent visual RAG studies leverage LVLMs to encode document images directly (Tanaka et al., 2025), enabling retrieval over visually rich documents. However, existing datasets such as ViDoRe (Wang et al., 2025) cover limited document types and often contain questions that do not truly require retrieval, and previous approaches typically lack dedicated training to adapt LVLMs for retrieval tasks.

VDocRAG (Tanaka et al., 2025) addresses these gaps with a dual-encoder retriever, where query tokens and document image features (processed by image encoder + projector) are fed into the same LVLM block to produce embeddings for similarity search. Its generator then uses the top-k retrieved images to produce answers. The model is built on Phi-3-Vision-128K-Instruct (4.2B parameters, image encoder + connector + projector + Phi-3 Mini LLM, 128K context length) and pre-trained with retrieval- and generation-oriented objectives (RCR, RCG) to align visual and textual features. OpenDocVQA [4], the accompanying dataset, provides open-domain and multi-hop questions, forming a comprehensive benchmark for visually rich document understanding.

2.3 Query Rewriting in Information Retrieval

Query rewriting is a common technique in information retrieval for reformulating user queries into semantically richer or more precise forms to improve retrieval performance. Existing approaches include rule-based methods, neural sequence-to-sequence models (Yu et al., 2020; Ma et al., 2023), and reinforcement learning strategies that optimize retrieval metrics (Ma et al., 2023). Recent work such as the Rewrite–Retrieve–Read framework demonstrates that rewriting can substantially improve dense retrievers by bridging the semantic gap between user queries and relevant documents (Ye et al., 2023; Kostric and Balog, 2024; Mo et al., 2023). However, most prior research focuses on text-only corpora, leaving open challenges for visually rich documents where key information may be embedded in layouts, figures, and tables.

3 Materials and Methods

3.1 Problem Definition

Let D denote a collection of image-centric documents (e.g., charts, tables, engineering drawings). We represent it as:

$$D = \{(Q_i, I_i)\}_{i=1}^{N}, \quad Q_i = \{q_{i1}, q_{i2}, \ldots, q_{ik}\},$$

where each image I_i may correspond to multiple associated queries, which are often ambiguous and underspecified. Our objective is

[1] https://www.paddleocr.ai/latest/en/version3.x/algorithm/PP-OCRv5/PP-OCRv5.html

[2] https://github.com/mindee/doctr

[3] https://github.com/JaidedAI/EasyOCR

[4] https://huggingface.co/datasets/NTT-hil-insight/OpenDocVQA-Corpus

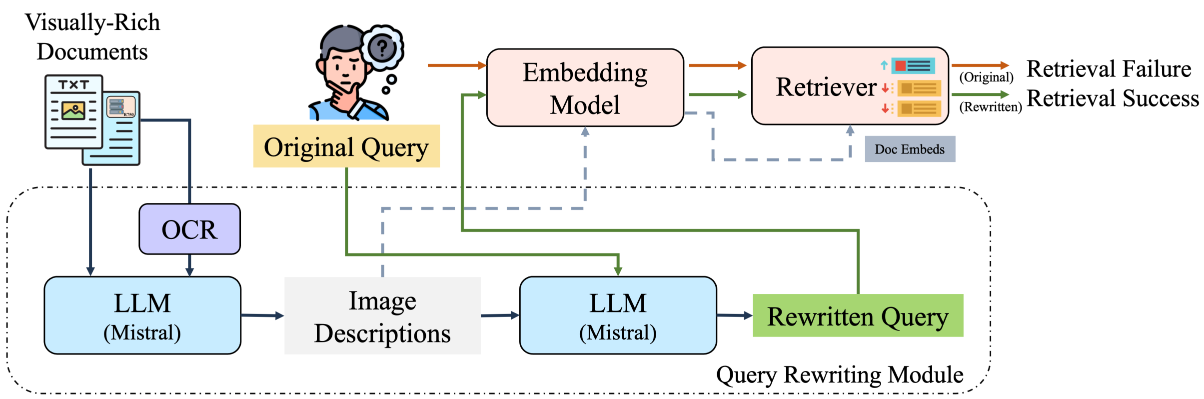


Figure 1: *Overview of the two-stage query rewriting framework.* Given an ambiguous query and a visually rich document, OCR text is extracted and summarized by a language model (Mistral) into an image description. The description, together with the original query, is used to generate a rewritten query that clarifies entities and avoids answer leakage. Both the original and rewritten queries are compared in the retriever to evaluate improvements in retrieval success.

to rewrite each query q_{ij} into a semantically complete and retrieval-friendly form $\widetilde{q}_{ij}$.

$$\widetilde{q}_{ij} = f_{\text{rewrite}}(q_{ij}) \tag{1}$$

The rewritten query $\widetilde{q}_{ij}$, generated through the two-stage process (Section 3.2), is constrained by a predefined system prompt (Section 3.3.1). Its objective is to retrieve the corresponding document more accurately.

3.2 Two-Stage Query Rewriting

The proposed two-stage framework (Figure 1) comprises image description generation and constrained query reformulation, detailed in the following subsections.

3.2.1 Image Description Generation

Each document image I_i is first processed by an OCR engine (PP-OCRv5; (Cui et al., 2025)) to extract the raw textual content $t_i = \text{OCR}(I_i)$. Since OCR outputs are often fragmented or incomplete (e.g. isolated labels or numbers), we employ Mistral-Small 3.2 (24B)[5] as the description generator f_{desc} to produce a context rich description d_i conditioned on both t_i and I_i:

$$d_i = f_{\text{desc}}(t_i, I_i) \tag{2}$$

The generated description supplements missing or implicit OCR details, providing essential context for the subsequent rewriting stage.

The choice of Mistral-Small 3.2 (24B) was validated through comparisons with lighter multimodal models (LLaVA-7B and Qwen2.5-VL-7B) in the dense retrieval configuration. Although the smaller models achieved moderate accuracy (68 Hits@1) with shorter and less coherent descriptions, the 24B variant generated richer and layout-aware outputs, yielding +8-9 higher Hits@1 and a favorable cost-performance balance.

3.2.2 Controlled Query Rewriting

In the second stage, the original query q_{ij} is rewritten into $\widetilde{q}_{ij}$ with the help of the image description d_i. The concatenated pair (q_{ij}, d_i) s fed into an LLM-based rewriting model f_{rewrite} (Mistral-Small 3.2 (24B)), guided by a structured prompt P and few-shot exemplars ξ (Section 2.3):

$$\widetilde{q}_{ij} = f_{\text{rewrite}}(q_{ij}, d_i \mid P, \xi) \tag{3}$$

This design allows the model to contextualize visual information via d_i and generate retrieval-friendly reformulations.

3.3 Prompt and Constraint Design

3.3.1 System Prompt

We design the rewriting prompt with explicit instructions that serve as hard constraints to ensure retrieval-oriented outputs. Specifically, the prompt requires that the rewritten query adhere to the following rules:

1. **Preserve interrogative form**: retain the question structure (e.g., "what," "how

[5] https://ollama.com/library/mistral-small3.2:24b

	# Queries	# Docs	Representative Visual Elements
Sales	135	25	Workflow and configuration diagrams; market analysis charts; wiring schematics; product dimension and application illustrations
Manufacturing	35	6	Process flow diagrams; Gantt charts; dimensional drawings; production statistics plots
Quality Control	52	9	Pareto and pie charts; Gantt charts; statistical performance plots
Technical	95	19	System layouts; architecture diagrams; measurement charts; circuit schematics
Others	40	7	Organizational and process flow diagrams
Total	357	66	—

Table 1: Domain-level statistics of the proprietary dataset containing 66 visually rich document images and 357 queries across five enterprise domains, each characterized by distinct visual elements common to industrial documentation.

many," "why") when the original query is interrogative.

2. **Avoid answer leakage**: exclude factual answers or numeric values appearing in the image text.

3. **Disambiguate references**: replace vague terms (e.g., "this chart," "the server") with concrete entities from d_i.

4. **Maintain source language**: keep the rewritten query in the same language as the input.

By enumerating these constraints in the system prompt, the model adheres to the intended query style and retrieval objectives.

Original Query	Rewritten Query
What kind of coating is applied to the machine surface?	What color of heat-resistant paint is applied on the surface of the ZX-200 industrial machine?
What is the efficiency improvement shown in the chart?	In the Q3 operations report, what is the percentage of efficiency improvement related to production output and cost reduction?
What is the memory specification of this server?	What is the memory configuration of the NovaEdge R720 server used in enterprise datacenter deployments?

Table 2: Query rewriting examples illustrating how ambiguous user questions are refined into precise, retrieval-oriented formulations.

3.3.2 Few-Shot Exemplars

To further guide model behavior, the prompt includes a few demonstration pairs of original and rewritten queries. Positive exemplars show effective reformulations where ambiguous queries are clarified with explicit entities or technical terms without leaking answers, while negative exemplars illustrate undesirable cases such as declarative rewrites, answer exposure, or language alteration. Together with the system prompt constraints (Section 3.3.1), these exemplars provide complementary supervision that steers f_{rewrite} toward generating well-formed, retrieval-oriented queries.

3.4 Post-hoc Validation

After rewriting, a lightweight validation step verifies compliance with the constraints in Section 3.3.1. This step ensures that each query preserves interrogative form, retains the original language, and avoids revealing factual answers or numeric values. Queries failing validation are replaced with the original input and logged with a status code, serving as a safeguard for overall quality and consistency.

3.5 Dataset

We evaluate the proposed method on a proprietary dataset provided by an industry partner, comprising 66 visually rich document images across five enterprise domains—sales, manufacturing, quality control, technical, and others. Each document contains layout-dependent visual structures such as charts and diagrams. A LVLM (Qwen3-VL-235B) was prompted to generate multiple natural-language queries per image, yielding 357 query-image pairs that simulate realistic but often underspecified informa-

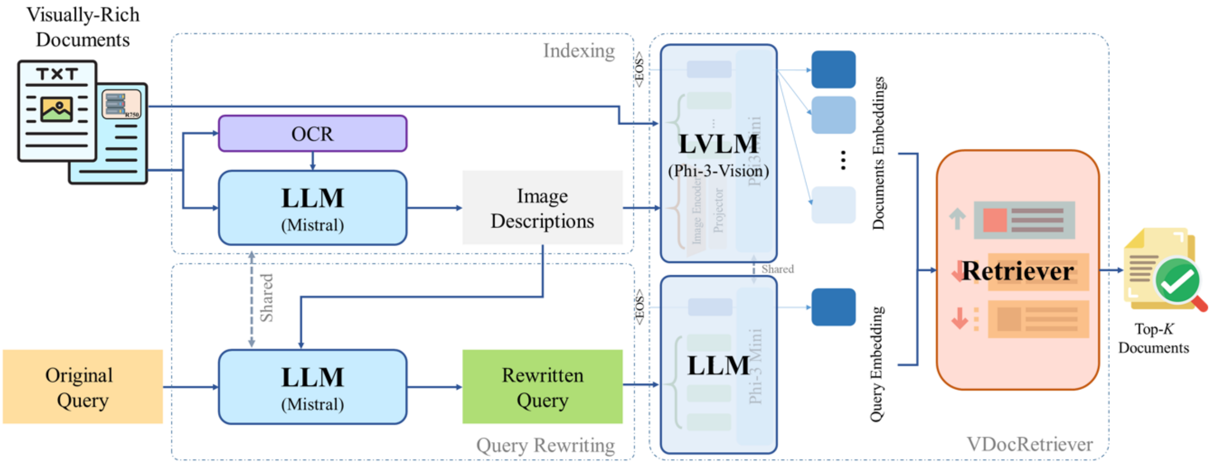


Figure 2: *Integration of the query rewriting module with VDocRetriever.* Each document is processed with OCR and a language model (Mistral) to generate enriched descriptions, which are combined with original queries for rewriting. Document and query embeddings are encoded by a vision—language model (Phi-3-Vision) to support multimodal retrieval. The red arrow marks VDocRetriever†, a variant that augments the document encoder with OCR-based image descriptions as additional textual context.

tion needs (Table 2). Each query q_{ij} is paired with its originating image I_i as the sole relevant item for retrieval.

To quantify query ambiguity, we manually classified all queries into three levels—clear (33.3%), partially underspecified (35.3%), and severely underspecified (31.4%)—based on the contextual information required for accurate retrieval. Although the dataset cannot be released due to confidentiality, detailed statistics (Table 1) and experimental results (Table 3) illustrate its diversity, the prevalence of ambiguous queries, and the effectiveness of the proposed framework.

4 Experiment Setup

4.1 Evaluation of Query Rewriting Across Retrieval Methods

We evaluate the proposed framework across three representative retrieval paradigms—dense, hybrid, and visual document retrieval (Figure 2)—covering neural, neural-lexical, and multimodal approaches. In each setting, rewritten queries replace the originals under identical conditions to isolate the effect of rewriting. Details of each retrieval configuration are provided in the following sections.

4.1.1 Dense Retrieval

For dense retrieval, we adopt BGE-M3 (Chen et al., 2024), a multilingual embedding model trained with contrastive objectives for retrieval tasks. Both queries and OCR-derived document descriptions are encoded into the same semantic space, and cosine similarity is used to rank document candidates. This text-only setup provides a strong baseline for evaluating whether query rewriting enhances semantic alignment between queries and OCR-based document representations.

4.1.2 Hybrid Retrieval

To leverage both semantic and lexical signals, we adopt a hybrid retrieval strategy combining BGE-M3 (Chen et al., 2024) and BM25 (Robertson and Zaragoza, 2009). Each rewritten query is simultaneously encoded by BGE-M3 for dense similarity matching and submitted to a BM25 index built from OCR-derived document text. BM25 first retrieves the top-k candidates; then both BM25 and cosine similarity scores are normalized to [0,1] and linearly combined ($0.6 \times$ BM25 $+ 0.4 \times$ BGE-M3), as tuned on validation data. This design prioritizes exact lexical matches while allowing semantic reranking, enabling controlled analysis of how rewriting affects both retrieval signals.

4.1.3 Visual Document Retrieval

We further evaluate VDocRetriever (Tanaka et al., 2025), a state-of-the-art system for visually rich document retrieval. Unlike dense or hybrid retrievers that rely solely on textual representations, VDocRetriever jointly encodes multimodal signals (layout, visual appearance, and OCR text) making it an ideal baseline for testing the robustness of query rewriting under multimodal retrieval.

Two configurations are considered: the original VDocRetriever, which jointly encodes queries and document images, and VDocRetriever†, which augments document embeddings with OCR-based descriptions as additional textual context. The latter allows us to examine whether explicit textual anchors further enhance cross-modal alignment when combined with query rewriting.

4.2 Evaluation Metric

Retrieval effectiveness is measured using the Hits@k metric, reported at $k = 1, 5, 10$. A query is counted as successful if its relevant document appears within the top-k retrieved results. Formally, for a set of queries $\{q_i\}_{i=1}^{N}$, Hits@k is defined as:

$$\text{Hits@k} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{rank}(d_i^*|q_i) \leq k] \quad (4)$$

where d_i^* denotes the ground-truth document for query q_i, and $\text{rank}(d_i^*|q_i)$ denotes the rank position of d_i^* returned by the retrieval system. $\mathbf{1}[\text{rank}(d_i^*|q_i) \leq k]$ is an indicator function that equals 1 if the condition is true (i.e., if the relevant document d_i^* for query q_i is ranked within the top-k results) and 0 otherwise.

Since each query in our dataset has a single relevant document, Hits@k directly reflects the ability of each retrieval configuration to surface the correct document near the top of the ranked list. Higher values (especially for small k) indicate better retrieval effectiveness.

5 Results & Discussion

Rewriting Models	Hits@1	Hits@5	Hits@10
Qwen 3 (4B)	56.3	75.9	79.6
Qwen 3 (14B)	57.4	76.6	79.6
Llama 3 (8B)	56.0	75.4	78.1
Mistral-Nemo (12B)	74.5	82.9	84.0
Mistral-Small 3.2 (24B)	**76.8**	**82.9**	**84.6**

Table 4: Performance comparison of different query rewriting models evaluated under the dense retrieval configuration (BGE-M3).

Across all retrieval configurations, query rewriting consistently improves retrieval effectiveness (Table 3). For the dense retriever (BGE-M3), Hits@1 increases from 57.4% to 76.8% (+33.8%), showing stronger semantic alignment between rewritten queries and OCR-based document embeddings. The hybrid retriever (BGE-M3 + BM25) exhibits a similar pattern (Hits@1 + 37.8 %), suggesting that rewriting introduces lexical cues that complement dense representations.

The most pronounced gains occur in multimodal retrieval. The baseline Hits@1 of VDocRetriever (21.0%) is considerably lower than that of dense or hybrid retrievers, reflecting the difficulty of aligning vague queries with image-based embeddings. Rewritten queries introduce explicit anchors—such as entity names, field labels, and technical terms—that facilitate cross-modal alignment, raising Hits@1 to 56.6% (+169.5%). With additional OCR-based image descriptions (VDocRetriever†), performance further improves to 79.3% Hits@1 and 97.8% Hits@10, approaching near-perfect retrieval. These results highlight the value of multimodal fusion, where textual anchors extracted from images mitigate ambiguity in visual representations and strengthen query—document alignment.

Beyond retrieval paradigms, we also analyzed the influence of the rewriting backbone (Table 4). Model capacity correlates with rewriting precision: smaller models such as Qwen 3 (4B/14B) and Llama 3 (8B) produced syntactically correct but semantically shallow rewrites, while Mistral-Nemo (12B) and Mistral-Small 3.2 (24B) generated more contextually grounded reformulations, achieving 74.5 and 76.8 Hits@1, respectively. The 24B model slightly outperformed the 12B variant while maintaining acceptable inference latency, making Mistral-Nemo (12B) a practical choice for cost-sensitive deployments, whereas Mistral-Small 3.2 (24B) remains preferable for high-precision retrieval.

Taken together, these findings reveal several key insights. First, query rewriting benefits both dense and hybrid retrieval, but has the greatest impact in multimodal settings. Second, the disproportionate gains observed for VDocRetriever highlight that query rewriting is most critical when retrieval relies heavily on visual or layout-based representations. Finally, the strong performance of VDocRe-

Target Document	Retrieval Method	Original Queries Hits@1	Original Queries Hits@5	Original Queries Hits@10	Rewritten Queries Hits@1	Rewritten Queries Hits@5	Rewritten Queries Hits@10
d_i	BGE-M3	57.4	76.8	79.6	76.8	82.9	84.6
d_i	BGE-M3+BM25	55.5	77.3	80.1	76.5	83.2	87.1
I_i	VDocRetriever	21.0	51.5	68.6	56.6	88.5	94.1
$I_i + d_i$	VDocRetriever†	29.4	52.4	64.2	79.3	93.8	97.8

Table 3: Retrieval performance with and without ablation study on the effect of query rewriting across different retrieval methods. Retrieval effectiveness is reported using Hits@k (%). † indicates the variant of VDocRetriever that incorporates the image description as additional context to improve retrieval precision. The column "Target Document" specifies the representation used as the retrieval target, such as document image embeddings I_i or OCR-based descriptions d_i.

triever† shows that combining rewriting with textualized visual context offers a powerful strategy for visually rich document retrieval. Overall, the results position query rewriting as a robust and versatile technique, capable of enhancing retrieval effectiveness across both text-centric and multimodal paradigms.

6 Conclusion

This paper presents a two-stage query rewriting framework for addressing underspecified queries in RAG systems over visually rich documents. By leveraging OCR-informed image descriptions and applying constrained reformulation, the framework produces retrieval-friendly queries that reduce ambiguity and improve alignment with document content. Experimental results demonstrate that query rewriting consistently enhances retrieval effectiveness across dense, hybrid, and visual document paradigms, with particularly strong benefits in visual document settings. Overall, the findings establish query rewriting as a robust and general strategy for RAG over visually rich documents, with promising potential for scaling to larger datasets and integration into end-to-end question answering pipelines.

References

Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani, Mohammed Ali, and Adam Jatowt. 2025. From retrieval to generation: Comparing different approaches. *arXiv preprint arXiv:2502.20245*.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. *Proceedings of the IEEE International Conference on Computer Vision*, pages 973–983.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2318–2335.

Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, Daoyu Wang, and Enhong Chen. 2025. A survey on knowledge-oriented retrieval-augmented generation. *arXiv preprint arXiv:2503.10677*.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *Proceedings - 2024 Conference on AI, Science, Engineering, and Technology, AIxSET 2024*, pages 166–169.

Ivica Kostric and Krisztian Balog. 2024. A surprisingly simple yet effective multi-query rewriting method for conversational passage retrieval. *SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2271–2275.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 5303–5315.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian Yun Nie. 2023. Convgqr: Generative query reformulation for conversational search. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:4998–5012.

Feifei Pan, Mustafa Canim, Michael Glass, Alfio Gliozzo, and James Hendler. 2022. End-to-end table question answering via retrieval-augmented generation. *arXiv preprint arXiv:2203.16714*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework. *Foundations and Trends in Information Retrieval*, 3:333–389.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. *Proceedings of the International Conference on Multimedia Information Processing and Retrieval, MIPR*, pages 155–161.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2025. Vdocrag: Retrieval-augmented generation over visually-rich documents. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24827–24837.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 20:1192–1200.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1:1933–1936.

基於微調開源大型語言模型的交通事故資訊蒐集代理人系統研究
The Study of a Traffic Accident Information Collection Agent System Based on Fine-tuned Open-Source Large Language Models

龔若齊 **Jo-Chi Kung** 張嘉惠 **Chia-Hui Chang**
z1a2x3s4c5d6v7f8b9g@gmail.com, chiahui@g.ncu.edu.tw

摘要

本研究提出了一套名爲「交通事故資訊蒐集代理人」（Collision Care Guide, CCG）的系統架構，專注於事故初期階段的結構化資訊蒐集。CCG 整合三大模組：問題生成、資訊擷取及事故重建，透過多輪對話引導使用者敘述事故細節並轉換爲結構化資料格式（TARF），同時生成可讀性敘述供核對。爲滿足成本效益、隱私保護及部署彈性需求，本研究比較開源 Llama 模型（3B/8B 參數，完整微調及 4-bit PEFT 方法）與商業基準 GPT-4o-mini 的效能表現。結果顯示，資訊擷取模組欄位準確率高於 0.94，JSON 語義相似度達 0.995；問題生成模組語義相似度介於 0.85-0.88，問題表達更加精煉。微調模型在對話品質與資訊擷取的 LLM 評估中均獲得 4 分以上（滿分 5 分），與商業基準差距小於 0.5 分。研究證實開源模型經微調後能逼近商業模型效能，且量化版本在資源受限場景中具備高效能與部署潛力。CCG 的設計填補了事故初期互動式資訊蒐集的技術空白，爲交通事故處理提供了高效且具成本優勢的解決方案。

Abstract

This study introduces the "Collision Care Guide" (CCG), a system designed to collect structured traffic accident information during the early stages of an incident. CCG integrates three core modules: question generation, information extraction, and accident reconstruction. Through multi-turn dialogues, users are guided to describe accident details, which are then transformed into a structured format (TARF), alongside readable narratives for verification. To address cost efficiency, privacy protection, and deployment flexibility, this study compares the performance of open-source Llama models (3B/8B parameters with full fine-tuning and 4-bit PEFT methods) against the commercial baseline GPT-4o-mini. Results show that the information extraction module achieves field accuracy above 0.94 and JSON semantic similarity of 0.995, while the question generation module attains semantic similarity between 0.85-0.88 with more concise expressions. Fine-tuned models scored 4 (out of 5) in dialogue quality and information extraction evaluations, with differences from the commercial baseline within 0.5 points. Findings confirm that fine-tuned open-source models can achieve performance comparable to commercial models, with quantized versions demonstrating high efficiency and deployment potential in resource-constrained scenarios. The CCG design bridges the technical gap in interactive information collection during the early stages of accidents, offering a cost-effective and efficient solution for traffic incident management.

關鍵字：交通事故資訊蒐集、大型語言模型、對話式 Agent、資訊擷取、模型微調

Keywords: LLM、Conversation Agent、Information Extraction、Finetuning

1 緒論

道路交通安全問題長期爲全球公共衛生與基礎設施治理的重大議題[1]。在台灣，年度交通事故總件數自 2019 年約 34 萬件攀升至 2024 年將近 40 萬件[2]，平均每日逾千件事故。事故初期的人工筆錄流程面臨多重挑戰：當事人在事故後受情緒與壓力影響，難以完整準確地有條理敘述事故經過；事故量的攀升亦造成執法人員的人力資源負擔，影響後續責任釐清的效率與公正。上述背景凸顯了交通事故資訊蒐集的需求，尤其是在高事故量與人工處理效率間的矛盾，爲自動化解決方案提供了切入點。

[1] https://www.who.int/publications/i/item/9789240087712
[2] https://ba.npa.gov.tw/statis/webMain.aspx?k=defjsp

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 71–79
November 20-22, 2025 ©2025 Association for Computational Linguistics

現有交通人工智慧研究多著重事故風險預測 (Zhou et al., 2020)、多模態重建與責任分析 (Wu et al., 2024)，法律科技領域聚焦判決預測與文本檢索 (Chalkidis et al., 2022)。然而，針對事故發生後初期之即時、互動式、結構化的資訊蒐集相對缺乏。現有技術主要假設事故事實已充分記錄，忽略了事故初期口語敘述的稀疏性及不確定性，這形成了前置事實蒐集的技術缺口。

爲填補此缺口，本研究提出基於大型語言模型的交通事故資訊蒐集代理人 Collision Care Guide（CCG），透過多輪對話引導提問與結構化模板，實現使用者口語敘述與結構化紀錄的雙向轉換。CCG 系統由三個協同模組構成：（1）問題生成模組：依據缺失資訊與前輪回覆動態生成聚焦提問；（2）資訊擷取模組：自當事人口語敘述中擷取並填入結構化 JSON 欄位，處理模糊與不確定表述；（3）事故重建模組：將完成之結構化紀錄重建爲條理清晰的敘述文本，以支援後續人員閱讀與理解。這些模組相互協作，形成完整的資訊蒐集流程，旨在提升事故初期事實蒐集的效率與準確性。

本研究基於Kung et al. (2024) 的研究，進一步考量部署成本、隱私與可控性需求，針對開源中小參數 Llama 模型（Llama 3.2 3B、Llama 3.1 8B）進行任務特化微調，並與商業基線 GPT-4o-mini 比較於兩項核心任務：資訊擷取與問題生成，同時評估多任務（Combined）訓練設定下之效能穩定性。測試集結果顯示，資訊擷取模組整體 JSON 語義相似度約 0.995，欄位完全準確度最高達 0.95；問題生成模組平均語義相似度達 0.85，提問長度平均差異約 20%，保留必要語義而減少冗餘同理心鋪陳。此外，4-bit 量化模型在保持高效能的同時，顯著降低了部署成本，驗證其在私有化部署中的潛力。

本文的主要貢獻如下：

1. 提出交通事故多輪對話資訊蒐集系統（CCG），能從當事人敘述中引導並擷取預先定義之關鍵資訊，實現結構化資料的雙向轉換，降低重複詢問與遺漏風險
2. 探索任務特化微調與多任務聯合訓練對兩核心任務（問題生成與資訊擷取）的穩定性與效能影響
3. 建立雙層評估（測試集指標與 LLM 評分），驗證小參數模型部署的成本效益與技術可行性

綜上所述，CCG 的設計不僅填補了事故初期資訊蒐集的技術空白，也通過實驗結果證實了開源模型的效能逼近商業模型的可行性，爲交通事故場景提供了一個高效且具成本優勢的解決方案。

2 相關研究

交通人工智慧技術在事故風險預測、事故重建與責任分析等領域取得了顯著進展。例如，Zhou et al. (2020) 等人提出的事故風險預測模型利用道路特徵與歷史事故數據進行風險評估，準確率達 90% 以上。Wu et al. (2024) 等人基於影像與感測器數據進行事故場景的重建，有效支持責任分析。然而，這些模型依賴完整的結構化輸入數據，難以應對事故初期口語敘述的稀疏性與不確定性。

法律科技研究主要集中於判決預測與文本檢索。例如，LexGLUE(Chalkidis et al., 2022) 基準系統能有效支持法律決策，但假設輸入事實已整理完備，難以處理事故初期的模糊敘述。保險系統則著重於理賠流程的自動化，依賴完整事故報告，缺乏即時處理口語敘述的能力。

開源模型微調研究顯示，針對特定領域任務的參數高效微調（PEFT）(Hu et al., 2021) 與量化技術 (Dettmers et al., 2023) 能在保持效能的同時降低部署成本。PEFT 技術通過調整少量模型參數，使模型能快速適應特定任務，顯著降低訓練成本。例如，在交通事故場景中，PEFT 技術能支持模型快速適應口語敘述的資訊擷取任務，提升事故初期資訊蒐集的效率。量化技術則通過使用低精度數據格式（如 4-bit 或 8-bit）來減少模型計算需求，保持效能的同時降低硬體資源消耗。Llama 系列模型 (Touvron et al., 2023) 在多任務學習與領域適應方面展現了潛力，爲私有化部署提供了高效且成本友善的替代方案。然而，針對事故初期資訊蒐集的特定場景，開源模型與商業模型的效能差異仍需系統性驗證。

綜上所述，現有研究在事故預測與分析方面已趨成熟，但事故初期的即時互動資訊蒐集仍缺乏系統化方案。同時，開源 LLM 在特定領域任務上的微調效能與部署可行性需要進一步驗證。因此，本研究聚焦於填補前置資料蒐集缺口，並探索成本效益與隱私友善的開源模型替代方案。

3 Collision Care Guide (CCG)

本章介紹交通事故資訊蒐集對話代理系統 Collision Care Guide (CCG)，其目標是在事故發生初期透過多輪互動提問取得事故關鍵資訊，並將資訊結構化，同時提供可讀性敘述供當事人核對。CCG 系統以模組化設計爲基礎，

整合三個核心模組：問題生成模組（Question Generation Module）、資訊擷取模組（Information Extraction Module）、事故重建模組（Accident Reconstruction Module），形成完整的資訊蒐集流程。

3.1 資料格式：TARF 與 QEF

TARF（Traffic Accident Record Format）是一個包含 18 個欄位的結構化資料格式，用於儲存事故相關資訊，包括基本情境、行爲路徑、環境條件、事件結果與動機用途等（如表 1 所示）。此格式確保事故資訊能以結構化方式進行存儲與檢索。

QEF（Question Explanation Format）則爲 TARF 中各欄位提供語義解釋與提問對齊說明，確保問題生成的語義精準性，減少術語歧義。例如，針對「事故發生地點」欄位，QEF 提供了具體的提問方式與語義參考。爲 TARF 各欄位提供語義解釋與提問對齊說明，確保問題生成的語義精準性，減少術語歧義。

Table 1: TARF 主要欄位與簡述

“事故發生日期”: 事故之具體日期
“事故發生時間”: 事故之具體時間
“事故發生地點”: 發生道路或地址
“對方駕駛交通工具”: 對方車種
“我方駕駛交通工具”: 我方車種
“我方行駛道路”: 我方所行經道路
“事發經過”: 簡述事故情節
“我方行進方向的號誌”: 相關號誌狀態
“當天天候”: 天氣情況
“道路狀況”: 施工／濕滑等狀態
“我方行車速度”: 事發時速度
“我方車輛損壞情形”: 我方車損
“我方傷勢”: 我方傷害
“對方車輛損壞情形”: 對方車損
“對方傷勢”: 對方傷害
“我方從哪裡出發”: 出發起點
“我方出發目的地”: 目的地
“我方出發目的是什麼”: 出發動機

3.2 模組化架構與多輪流程

CCG 系統採用模組化設計，由三個模組協同運作，形成完整的資訊蒐集與驗證流程。整體流程採用缺失導向的迭代策略（如圖 1 所示），具體包括以下步驟：**1.** 初始詢問：系統通過開放式問題獲取事故概要。**2.** 迭代循環：系統檢測 TARF 中的缺失欄位，並動態生成聚焦問題以補全資訊。**3.** 結構化更新：根據使用者回答更新 TARF 欄位，直至所有欄位填寫完成或達到最大輪數限制（20 輪）。**4.** 敘述重建：流程最後階段將結構化紀錄重建爲自然語言敘述供使用者核對，必要時進行局部修正。此設計確保資訊完整性與一致性，並形成可驗證的雙向轉換閉環。

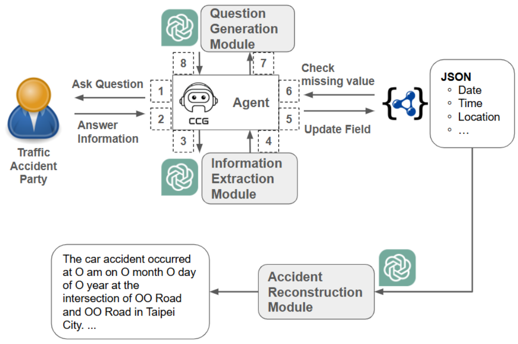


Figure 1: CCG 架構與多輪採集流程

3.3 問題生成模組

問題生成模組負責根據 TARF 填寫狀態與使用者回覆動態生成適當的回應與後續問題。該模組整合以下三項功能：在互動回覆方面，模組能對偏題回答進行專業引導，並對正確回覆給予正向回饋，從而維持良好的互動氛圍。動態檢查使模組能主動識別 TARF 中的未填欄位，結合 QEF 語義解釋生成自然且具體的提問，確保資訊蒐集的完整性。此外，語境維持透過參考前次對話內容（表 2），有效避免重複提問並保持語義連貫性。

例如，當 TARF 中「事故發生地點」欄位缺失時，模組會生成如下問題：「請問事故發生的具體地點是在哪裡？」

Table 2: 問題生成模組提示詞摘要（藍色文字表示每輪對話中動態替換的欄位資訊）

Prompt
作爲車禍事故敘述助理，主要任務包括：
1. 根據 [上一個問題] 和 [當事人回答] 給予適當回應
2. 根據 [下一個欄位] 詢問下一個問題
若回答不相關則重新引導；對焦慮當事人給予鼓勵
輸入參數：
- [上一個問題]：{previous_question}
- [當事人回答]：{user_response}
- [下一個問題]：{current_question}
- [問題解釋]：{qef_attributes}

3.4 資訊擷取模組

資訊擷取模組負責將使用者的自然語言回覆轉換爲結構化的 TARF 資料。該模組依據上下文理解回覆的語義脈絡，並遵循以下原則：（如表 3）1. 僅處理與當前問題直接相關的欄位，不進行跨欄推測。2. 面對「不記得」、「不知道」等回覆時，將欄位標記爲「未知」。3. 確保填入資訊忠實於原始回答內容。

例如，當使用者回答「高雄市楠梓區左楠路機車道」時，模組會更新 TARF 中「我方行駛

道路」欄位爲該內容。

Table 3: 資訊擷取模組提示詞摘要（藍色文字表示每輪對話中動態替換的欄位資訊）

Prompt
專業事件資訊擷取助理，從 [當事人回答] 中擷取資訊 並填入 [JSON 格式] 對應欄位 執行規則： • 僅處理與當前問題直接相關的 JSON 欄位 • 「不記得」、「不知道」等回應填入「未知」 • 確保填入資訊忠實於原始回答 輸入參數： - [JSON 格式]：{current_tarf} - [問題]：{previous_question} - [當事人回答]：{user_response}

3.5 事故重建模組

事故重建模組將結構化的 TARF 資料逆向轉換爲自然語言敘述，以驗證系統的雙向語義保真度（如圖 2所示）。此模組的設計體現了雙重資料格式的價值：結構化 **TARF 格式**：適合自動化處理與快速檢索的應用場景，如保險理賠處理與法律文件生成。自然語言敘述：更適合人員閱讀理解的情境，包括事故報告撰寫與當事人陳述確認。

例如，根據 TARF 中的資料，模組生成如下敘述：「事故發生於 2024 年 7 月 15 日早上 8:30，地點爲高雄市楠梓區左楠路。當時天氣晴朗，路面乾燥，事故涉及我方機車與對方轎車。機車行駛速度約 40 公里每小時，事故導致我方車輛右側損壞。」

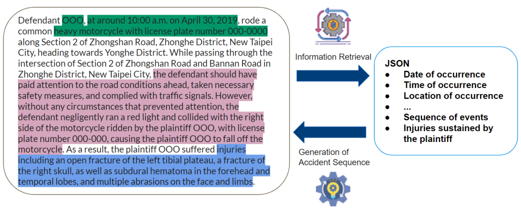


Figure 2: 資訊擷取與事故重建之雙向關係

Table 4: 事故重建模組提示詞摘要（藍色文字表示動態替換的欄位資訊）

Prompt
車禍諮詢專家，根據 [JSON 格式] 中的事實， 用敘述方式重述整個車禍經過 執行規則： • 僅描述 JSON 中提供的車禍相關事實 • 不包含其他無關或未提供的資訊 • 組織成連貫且邏輯清晰的事故敘述 輸入參數： - [JSON 格式]：{final_tarf}

4 模型訓練

本章驗證任務特化微調後之開源模型（Llama 3.2 3B，Llama 3.1 8B）在 CCG 系統的兩核心任務（問題生成、資訊擷取）中，分別進行單任務訓練與多任務聯合（Combined）設定，並透過測試集與 LLM 自動評估指標，檢驗其效能是否能逼近商業基準（GPT-4o-mini）。此外，本研究評估 4-bit 量化參數微調（PEFT）技術在成本、隱私與部署彈性場景中的替代可行性。

4.1 模型選擇與實驗環境

Llama 模型憑藉其在語義理解、生成能力及開源特性，成爲本研究的首選。相比商業模型（如 GPT-4o-mini），開源模型具有更高的成本效益及部署靈活性。爲進一步降低記憶體需求，採用 PEFT 技術（LoRA）進行 4-bit 量化，確保模型在資源受限場景中的實際應用能力。

所有訓練實驗均於 Nvidia RTX A6000 GPU 上執行，該 GPU 提供 48GB 的顯存，能有效支持大規模模型的微調，並顯著縮短訓練時間。整體訓練流程採用 Unsloth 框架[3]，該框架通過記憶體優化技術，顯著降低 GPU 記憶體占用並提升訓練效率。

4.2 訓練資料準備

4.2.1 資料格式設計

爲符合 CCG 系統的推論流程，本研究將對話狀態、缺失欄位提示、問題語境與生成約束整合爲單一 Instruction，並將輸出限定爲目標（下一個問題或更新後的 TARF）。此格式設計有助於模型專注於格式一致性與語義對齊，並便於單任務與聯合訓練模型共享資料結構。

每筆資料包括兩個主要部分：Instruction 與 Output。表 5 與表 6 分別展示了問題生成與資訊擷取的訓練資料示例。

Table 5: 問題生成訓練資料示例（節錄）

Instruction
作爲事故敘述助理，依上下文提出下一則問題。 上一問題：請簡述事故。回答：…… 欄位缺失：我方行駛道路； 欄位說明：事故時我方行駛的具體道路名稱。
Output
請問事故發生時您行駛的道路名稱是什麼？

4.2.2 標籤平衡策略

原始判決文本偏重法律裁判目的，導致 TARF 18 個欄位中僅約 8-10 個被明確提

[3] https://unsloth.ai/

Table 6: 資訊擷取訓練資料示例（節錄）

Instruction
依據回覆更新 JSON。 現有 JSON：{事故發生日期: ...}； 問題：您行駛的道路？ 回覆：高雄市楠梓區左楠路機車道。
Output
{事故發生日期: 民國 108 年 4 月 2 日, 事故發生時間: 07:28, 我方行駛道路: 高雄市楠梓區左楠路機車道, ...}

及。爲避免模型學得過於保守的策略，本研究對訓練資料進行標籤擴增，將資訊分爲三類：normal（準確資訊值）、unknown（明確未知）、other（模糊/無法解析）。經擴增後，訓練集的標籤分布爲 160:198:20（約 42.3%/52.4%/5.3%），測試集的分布爲 39:46:3（約 44.3%/52.3%/3.4%），確保具體值與未知標記的決策邊界更加平衡。

4.2.3 多樣性與品質控制

爲提升模型的泛化能力，本研究從敘述密度（詳細/中等/精簡）與回答風格（五類語用態度）生成多樣化訓練語料。敘述密度分爲：詳細型：涵蓋大部分 TARF 欄位，內容完整，類似完整事故報告、中等型：僅包含主要事故資訊，省略部分次要細節、精簡型：僅敘述核心事故事實，形式精煉。

回答風格包含五類：驚慌失措型（語序跳躍、重複）、冷靜理性型（邏輯線性、資訊密集）、防禦戒備型（對責任相關細節保留）、創傷恍惚型（不確定詞頻繁、時間順序模糊）、急躁不耐型（回答簡短、易省略修飾）。此考量真實情形中資訊不完整的場景，同時反映不同當事人的敘述習慣和記憶能力差異，旨在使訓練資料能夠反映真實世界中當事人的多樣化回答模式。

資料生成流程基於判決書樣式與 TARF 欄位模板，運用多個大型語言模型（Claude 4 Sonnet、Gemini-2.5-pro、GPT-4.1）產出三種事故敘述密度初稿，設定特定風格的當事人角色與基準模型 CCG 進行對話擴展並植入標籤，經人工審核格式合法後形成訓練/測試資料。

4.2.4 資料統計

最終資料集包含：訓練集 40 篇對話（資訊擷取 378 筆樣本、問題生成 338 筆樣本）與測試集 10 篇對話（資訊擷取 88 筆樣本、問題生成 78 筆樣本），資料分布保持與實際多輪流程相近，覆蓋全部密度與風格組合。

4.3 模型選擇與訓練配置

本研究比較 3B 與 8B 兩種模型規模，並採用完整微調（Full Fine-Tuning）與 4-bit 量化 LoRA（PEFT）技術進行對照。具體配置如表 7 所示。

Table 7: 模型訓練配置

名稱	版本	訓練方法	量化
Llama_3B_4bit	3.2 3B	PEFT (LoRA)	4-bit
Llama_3B	3.2 3B	Full FT	None
Llama_8B_4bit	3.1 8B	PEFT (LoRA)	4-bit
Llama_8B	3.1 8B	Full FT	None

5 模型效能評估結果

本章節旨在比較微調模型與基準模型 GPT-4o-mini 在資訊擷取與問題生成任務中的效能表現。評估過程採用 (Kung et al., 2024) 所蒐集的 754 筆對話資料，並透過雙評估器（Gemini-2.0 與 GPT-4o）進行交叉評分，以檢驗模型在對話品質及資訊擷取整體品質上的一致性與穩健性。

5.1 測試集驗證概述

本研究主要聚焦於兩項核心任務：資訊擷取與問題生成。在資訊擷取任務中，模型效能以欄位層級的精確性及語義保真度進行評估，並透過整體 JSON 一致性來衡量模型跨欄位的語義與結構表現。在問題生成任務中，則著重於模型生成問題的語義覆蓋率及表達精煉程度。此外，爲驗證多任務訓練的成效，研究進一步探討共享語義表示的穩定性是否得以維持或提升。

5.2 資訊擷取效能分析

爲評估模型效能，本研究採用六項指標：完全準確度（**Exact Accuracy**）衡量欄位輸出是否與基準一致；語義相似度（**Semantic Similarity**）以 Sentence Transformer 計算嵌入向量餘弦相似度，範圍 [-1,1]，越接近 1 語義越一致；高語義準確度（**High Semantic Accuracy**）爲語義相似度高於 0.8 的有效欄位比例；未知／空值不匹配率（**Unknown / Empty Mismatch Rate**）檢驗模型對不確定資訊的決策一致性；整體 **JSON** 相似度（**Overall JSON Similarity**）評估跨欄位語義與結構的一致性。

上述指標核心公式如下：

$$\text{Accuracy}_j = \frac{1}{|\mathcal{V}_j|} \sum_{i \in \mathcal{V}_j} \mathbf{1}\left(o_{i,j}^{(f)} = o_{i,j}^{(b)}\right)$$

Table 8: 資訊擷取模組測試結果（欄位層級彙總指標）

模型	Exact Accuracy	High Semantic Accuracy	Unknown Mismatch	Empty Mismatch	Overall JSON Similarity
3B	0.9508	0.9654	0.0139	0.0183	0.9946
3B_4bit	0.9508	0.9694	0.0139	0.0189	0.9948
8B	**0.9571**	0.9621	0.0145	0.0152	**0.9957**
8B_4bit	0.9539	0.9627	0.0139	0.0158	0.9942
Combined_3B	0.9489	0.9673	0.0170	0.0227	0.9942
Combined_3B_4bit	0.9470	0.9618	0.0158	0.0202	0.9937
Combined_8B	0.9558	**0.9702**	0.0139	0.0177	0.9954
Combined_8B_4bit	0.9558	0.9701	**0.0107**	**0.0145**	0.9954

$$\text{Unknown\_Mismatch}_j = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big(u_{i,j}^{(f)} \neq u_{i,j}^{(b)}\big)$$

其中，N 爲測試樣本總數（88）；$\mathcal{V}_j$ 爲欄位 j 之有效樣本集合；$o_{i,j}^{(f)}$、$o_{i,j}^{(b)}$ 分別爲微調與基準模型輸出；$\mathbf{1}(\cdot)$ 爲指示函數。

表 8 彙總了模型在資訊擷取任務中的效能指標。結果顯示，8B 模型在完全準確度（0.9571）及 JSON 相似度（0.9957）方面表現最佳，展現卓越的語義與結構一致性。4-bit 量化版本的 Combined_8B_4bit 模型在未知（0.0107）與空值（0.0145）不匹配率上最低，證明量化技術與多任務訓練未影響穩定性。相比之下，8B 模型效能略高於 3B 模型，而量化版本顯著降低計算資源需求，適合資源受限場景。

表 9 顯示，Combined_8B_4bit 模型大多數欄位的完全準確度超過 0.95，標準化欄位（如車種、天候）語義相似度達 0.985 至 1.000。長敘述欄位（如「事發經過」）的完全準確度較低（0.898），但語義相似度達 0.995，顯示模型能有效掌握長文本語義。道路狀況與速度欄位因表達多樣性導致未知不匹配率略高（≥ 0.034），但仍在可接受範圍。

綜上，微調模型在結構化欄位填寫與語義重現上接近商業基準，量化技術未對主要指標造成影響，多任務設定進一步提升語義泛化能力與穩定性。

Table 9: Combined_8B_4bit 各欄位指標

欄位	完全準確度	語義相似度	未知不匹配率
事故發生日期	0.955	0.982	0.011
事故發生時間	0.955	0.982	0.000
事故發生地點	0.955	0.984	0.011
對方駕駛交通工具	0.989	0.994	0.000
我方駕駛交通工具	0.977	1.000	0.000
我方行駛道路	0.977	0.996	0.000
事發經過	0.898	0.995	0.000
我方行進方向的號誌	0.966	1.000	0.011
當天天候	0.989	1.000	0.011
道路狀況	0.932	0.973	0.034
我方行車速度	0.955	0.910	0.034
我方車輛損壞情形	0.943	0.988	0.000
我方傷勢	0.932	0.992	0.000
對方車輛損壞情形	0.955	0.984	0.000
對方傷勢	0.932	0.965	0.023
我方從哪裡出發	0.989	1.000	0.000
我方出發目的地	0.977	0.985	0.023
我方出發目的是什麼	0.932	0.987	0.034

5.3 問題生成效能評估

本研究針對問題生成任務進行了全面的性能評估，測試集包含 78 筆樣本，主要分析模型在語義覆蓋及表達簡潔性方面的表現。評估指標包括：高語義準確度（語義相似度閾值 0.8 的比例）、中等語義準確度（語義相似度閾值 0.6 的比例）、平均語義相似度（Avg Semantic Similarity），以及長度相似度（Avg Length Similarity）。

長度相似度（**Length Similarity**）用於衡量微調模型生成問題的長度與基準模型生成問題長度的相似程度，反映問題表達的完整性與精煉性。該指標的計算公式如下：

$$\text{AvgLengthSim} = \frac{1}{N}\sum_{i=1}^{N} \frac{\min(|q_i^{(f)}|, |q_i^{(b)}|)}{\max(|q_i^{(f)}|, |q_i^{(b)}|)}$$

其中 $q_i^{(f)}$ 和 $q_i^{(b)}$ 分別代表微調模型與基準模型生成的第 i 個問題的長度，N 爲樣本總數。

表 10 彙總了各模型在問題生成任務中的測試結果。所有模型的平均語義相似度均達到 0.8323 以上，其中 Combined_8B_4bit 在高語義準確度（0.7949）及平均語義相似度（0.8841）方面表現最佳，而 Combined_8B 和其量化版本在中等語義準確度上均達到 0.9872，展現了近乎完整的語義覆蓋能力。此外，所有模型的長度相似度集中於 0.78 至 0.80，顯示微調模型生成的問題平均比基準模型短約 20%，但仍保留了核心語義，呈現更直接且精煉的提問風格。

以下以道路狀況欄位生成問題爲例，展示基準模型與微調模型的語言特徵差異：

Table 10: 問題生成模組測試結果

模型	High Semantic Accuracy	Avg Semantic Similarity	Medium Semantic Accuracy	Avg Length Similarity
3B	0.6538	0.8486	0.9103	0.7919
3B_4bit	0.6667	0.8323	0.8974	0.7848
8B	0.7179	0.8601	0.9487	0.7779
8B_4bit	0.6923	0.8545	0.9487	0.7850
Combined_3B	0.6667	0.8508	0.9359	0.7876
Combined_3B_4bit	0.7308	0.8671	0.9615	**0.7990**
Combined_8B	0.7564	0.8761	0.9872	0.7822
Combined_8B_4bit	**0.7949**	**0.8841**	**0.9872**	0.7939

- **GPT-4o-mini**：「理解您的情況，燈號有時可能會讓人感到模糊。那麼，當時您所行駛的道路狀況如何？例如，那條路上是否有施工、破損或者是路面濕滑？」
- **Llama-3B**：「請您回想一下當時行駛的道路狀況如何？例如，有沒有施工、路面是否濕滑或有其他障礙？」

基準模型的問題通常包含情境鋪陳及同理語句，旨在增加對話的自然性與情感連結；而微調模型則省略冗長的轉接語，直接聚焦於核心語義，符合任務功能性需求。這種語言特徵的差異表明微調模型在提升表達簡潔性的同時，仍能維持語義完整性。

5.4 對話品質綜合評估

本節針對 754 筆對話進行雙評估器交叉驗證，分別使用 Gemini-2.0 與 GPT-4o 評估對話品質，涵蓋三項核心指標。評估方法參考 GPTScore (Fu et al., 2023) 與 G-Eval (Liu et al., 2023) 等研究，採用統一提示模板進行綜合評分。評分採用 5 分制量表，指標定義如下：流暢性（**Fluency**）評估系統生成回應的語言自然度與流暢性；關聯性（**Relevance**）衡量回應內容與用户描述或問題的相關性；連貫性（**Coherence**）檢驗整體對話流程的邏輯一致性。

表 11 彙總了各模型在流暢性、關聯性、連貫性及整體評分的表現。結果顯示，Gemini-2.0 評估中，微調後的 Llama-8B 及其量化版本在整體評分上達到 4.74—4.76，與基準 GPT-4o-mini（4.65±0.60）僅有微小差距。相比之下，GPT-4o 的評分普遍低於 Gemini-2.0，下降幅度約 0.4—0.5 分，顯示 GPT-4o 採用了更嚴格的評估標準。

在流暢性（Fluency）方面，所有模型均達到高分，其中 GPT-4o-mini 在 Gemini-2.0 評估中得分爲 5.00±0.06，展現了極高的語言自然度。然而，微調後的 Llama-8B 及其量化版本在關聯性（Relevance）與連貫性（Coherence）指標上的表現更爲穩定，整體評分達到 4.74—4.76，接近基準商業模型 GPT-4o-mini。

相比之下，GPT-4o 評估結果顯示微調模型的整體得分略低，主要體現在關聯性與連貫性指標上。例如，Llama-8B 的 GPT-4o 評分爲 4.28±0.69，低於其在 Gemini-2.0 中的 4.76±0.62。這表明，GPT-4o 更加注重對話回應的語義深度與邏輯一致性，導致評分標準更爲嚴苛。

整體而言，微調後的 Llama 系列模型在對話品質評估中表現穩定，尤其在 Gemini-2.0 評估中接近商業模型的效能，證實了開源模型的潛力及量化技術的實務可行性。

5.5 資訊擷取能力綜合分析

本節針對同批對話資料進行資訊擷取能力的全面評估，涵蓋三項核心指標，均採用 5 分量表進行評分：事實一致性（**Fact Consistency**）檢驗 JSON 格式中提取的資訊是否準確反映對話內容；資訊完整性（**Information Completeness**）評估 JSON 是否涵蓋當事人描述的所有必要事故要素；描述合理性（**Description Reasonability**）判斷生成的資訊描述是否合乎邏輯、客觀且對未提及的資訊正確標記爲未知。

表 12 彙總了各模型在資訊擷取任務中的表現，涵蓋事實一致性、資訊完整性、描述合理性及整體評分。Gemini-2.0 評估結果顯示，所有模型的整體得分均達 4.93 以上，其中 Llama-8B-4bit 模型以 4.97 的得分表現最佳，與基準 GPT-4o-mini（4.96±0.20）差距極小，顯示其卓越的資訊擷取能力。

相比之下，GPT-4o 評估標準更爲嚴苛，模型得分略低於 Gemini-2.0，但 Llama-8B 系列模型仍保持穩定表現，整體得分（4.84—4.86）略高於基準 GPT-4o-mini（4.82±0.40）。此外，量化技術未對模型性能產生顯著影響，量化版本（4-bit）與未量化版本的得分差距不超過 0.03，證實其在資源受限場景中的部署價值。

Table 11: 對話品質評估結果（流暢性、關聯性、連貫性）

模型	Gemini-2.0				GPT-4o			
	Fluency	Relevance	Coherence	Overall	Fluency	Relevance	Coherence	Overall
GPT-4o-mini	5.00±0.06	4.64±0.61	4.63±0.62	4.65±0.60	4.98±0.14	4.50±0.65	4.51±0.62	4.64±0.47
Llama-3B	4.74±0.49	4.68±0.66	4.63±0.73	4.65±0.71	4.54±0.51	4.14±0.75	3.96±0.87	4.14±0.74
Llama-3B-4bit	4.72±0.49	4.67±0.63	4.61±0.71	4.63±0.68	4.51±0.52	4.19±0.76	3.99±0.90	4.20±0.74
Llama-8B	4.85±0.38	4.77±0.64	4.75±0.66	4.76±0.62	4.66±0.47	4.28±0.70	4.14±0.82	4.28±0.69
Llama-8B-4bit	4.82±0.40	4.76±0.61	4.73±0.65	4.74±0.62	4.54±0.50	4.11±0.72	3.94±0.83	4.12±0.71
Llama-3B (Combined)	4.73±0.48	4.67±0.67	4.62±0.73	4.63±0.70	4.54±0.51	4.15±0.73	3.96±0.85	4.15±0.73
Llama-3B-4bit (Combined)	4.68±0.52	4.62±0.68	4.54±0.81	4.57±0.77	4.49±0.51	4.06±0.81	3.86±0.92	4.06±0.79
Llama-8B (Combined)	4.84±0.38	4.76±0.66	4.75±0.66	4.76±0.63	4.60±0.49	4.20±0.72	4.04±0.83	4.19±0.72
Llama-8B-4bit (Combined)	4.85±0.36	4.77±0.65	4.76±0.64	4.76±0.63	4.63±0.48	4.24±0.72	4.09±0.83	4.24±0.72

Table 12: 資訊擷取評估結果（事實一致性、資訊完整性、描述合理性）

模型	Gemini-2.0				GPT-4o			
	Consistency	Completeness	Reasonability	Overall	Consistency	Completeness	Reasonability	Overall
GPT-4o-mini	4.96±0.20	4.94±0.24	4.98±0.14	4.96±0.20	4.70±0.50	4.80±0.41	4.98±0.17	4.82±0.40
Llama-3B	4.93±0.25	4.95±0.22	4.95±0.23	4.94±0.23	4.72±0.57	4.69±0.50	4.85±0.46	4.75±0.52
Llama-3B-4bit	4.92±0.33	4.95±0.24	4.94±0.29	4.93±0.30	4.73±0.57	4.67±0.53	4.86±0.45	4.75±0.52
Llama-8B	4.94±0.26	4.96±0.19	4.96±0.21	4.95±0.23	4.82±0.46	4.80±0.43	4.90±0.39	4.83±0.45
Llama-8B-4bit	4.95±0.22	4.97±0.18	4.97±0.18	4.97±0.19	4.86±0.42	4.81±0.42	4.92±0.33	4.86±0.40
Llama-3B (Combined)	4.94±0.28	4.94±0.27	4.94±0.27	4.94±0.27	4.76±0.55	4.68±0.53	4.88±0.44	4.78±0.51
Llama-3B-4bit (Combined)	4.94±0.30	4.95±0.28	4.96±0.27	4.96±0.27	4.74±0.58	4.71±0.53	4.86±0.47	4.76±0.53
Llama-8B (Combined)	4.95±0.21	4.96±0.19	4.96±0.20	4.96±0.20	4.84±0.46	4.81±0.41	4.90±0.39	4.84±0.45
Llama-8B-4bit (Combined)	4.95±0.22	4.96±0.19	4.96±0.20	4.96±0.21	4.84±0.46	4.81±0.41	4.90±0.39	4.84±0.44

5.6 結果分析與結論

本研究通過測試集比較及 LLM 自動評估，驗證了 Llama 微調模型在交通事故對話任務中的效能。結果顯示：

綜合測試集與 LLM 自動評估結果：1) 資訊擷取能力：微調模型生成的 JSON 語義相似度達 0.995，欄位準確度超過 94%，量化模型在資源受限環境中保持穩定性，準確度達952) 問題生成能力：平均語義相似度約 0.85，最佳高語義準確度達 0.7949，問題更精煉但語義完整性未受影響。3) **LLM** 評估結果：多任務訓練未稀釋模型品質，性能與單任務微調相當，部分指標略有提升。

微調模型在特定任務上效能卓越，但對突發場景的適應性仍不及通用模型。本研究證實了微調開源模型的高效性與成本效益，並展現其在資源受限應用場景中的潛力。

6 研究限制

儘管測試集與 LLM 評估結果表明 CCG 微調模型在語義準確性與欄位對齊方面表現出色，其效能及適用性仍受到以下限制的影響。

資料覆蓋有限：訓練資料主要來自判決書文本，細節欄位僅涵蓋約 8-10/18 欄位，影響模型在特定場景的泛化能力。代理生成資料部分降低偏差，但無法完全模擬壓力情境下的語用變異，可能導致回覆不連貫或斷裂。

功能範疇侷限：CG 系統僅限於結構化資訊蒐集，不涉及法律推理或裁量功能，需明確界定用途以避免誤用或誤解。

適用範圍侷限：模型僅適用於繁體中文及台灣法規環境，尚未驗證跨語言或異質法規體系的效能。實務部署中需考量隱私保護與法規遵循，探索技術與法律框架整合以確保合規與安全性。

7 結論

本研究提出 CCG 系統架構，整合問題生成、資訊擷取及事故重建三大模組，專注於交通事故初期的結構化資訊蒐集，爲警政初步紀錄提供技術支撐。

基於 Llama 模型的任務特化微調，系統性能接近商業模型：資訊擷取欄位準確率達 89%，生成 JSON 語義相似度 0.995；問題生成語義相似度 0.85–0.88，提問精煉效果提升 20%。多任務聯合訓練維持語義穩定性，4-bit 量化版本主要指標保持 95% 以上一致性，證實系統適合低資源環境的私有化部署。

不同於聚焦事前預測或事後分析的研究，CCG 專注於「事故後第一時間」的互動式資料收集，爲交通事故前期事實蒐集提供創新解決方案。研究表明，結構化設計與多任務微調使開源模型在專業領域逼近商業模型，展現高部署靈活性與成本效益，並爲 AI 跨領域應用奠定基礎。

References

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Jo-Chi Kung, Chia-Hui Chang, Huai-Hsuan Huang, and Kuo-Chun Chien. 2024. A narrative assistant for traffic accidents based on large language models (llm). In *Legal Knowledge and Information Systems*, pages 84–94. IOS Press.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

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. 2023. Llama: Open and efficient foundation language models.

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

Zhengyang Zhou, Yang Wang, Xike Xie, Lianliang Chen, and Hengchang Liu. 2020. Riskoracle: A minute-level citywide traffic accident forecasting framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1258–1265.

Automatic Generation of Corpus-Based Exercises Using Generative AI

Adrian Jan Zasina
Institute of Czech and Deaf Studies
Charles University
Prague, Czech Republic
adrian.zasina@ff.cuni.cz

Abstract

This study explores the automatic generation of corpus-based language exercises using a generative AI model Corpus Linguist. It focuses on the interaction between the language model and corpus data, detailing a workflow in which collocation and translation patterns are extracted from a tagged corpus and structured prompts are constructed to guide the model in producing sentence-level exercises. The generated exercises reveal both the potential and the current limitations of AI-driven approaches. Challenges include inconsistency in corpus data use, and choosing appropriate translation equivalents. These observations highlight the necessity of careful design and critical evaluation when integrating generative models with corpus-based language materials. By analysing these processes from a computational linguistics perspective, this study contributes to understanding how generative AI can interact with structured linguistic data, informing future applications in automated language resources.

Keywords: AI, corpus, corpus-based exercises, DDL

1 Introduction

Since November 2022, when ChatGPT from OpenAI was released, new language models using generative artificial intelligence (GenAI) have appeared. These are no longer simple chatbots but more advanced tools that allow users not only to engage in conversation but also to create images and videos, and perform data analysis. Thanks to this development, it has become possible, for the very first time, to link corpus data with GenAI to create human-like queries for corpora, instead of relying on traditional queries in Corpus Query Language (CQL), which require professional knowledge of CQL syntax. Combining corpus methods with GenAI opens new possibilities in language analysis, enabling us to evaluate how GPT models interact with corpus data.

One reason for this synergy is the integration of corpus data and GenAI with data-driven learning (DDL), which relies on authentic data in language learning (Johns, 1991). In the early 1990s, DDL faced challenges due to the lack of user-friendly tools that could attract learners. Some studies (Vyatkina, 2020, pp. 362–363; Zasina, 2022, p. 126) highlight that learners benefit from corpus activities, however, the tools available at the time were often too complex. The advent of ChatGPT has enabled fast communication and brought substantial computational capacities. Finally, the creation of Corpus Linguist GPT model (Milička & Machálek, 2024) in 2024 has made it feasible to use corpus data in a user-friendly environment. Some corpus linguists are aware of the limitations of existing corpus interfaces, therefore, new studies (Cheung & Crosthwaite, 2025) combining these two sources are beginning to appear.

However, there is little evidence on the interaction between language models and corpus data in terms of DDL practice (cf. Zasina, 2025). Therefore, this study attempts to explore the automatic generation of corpus-based exercises for learners using a GenAI model. First, attention is paid to prompt crafting along with the co-star framework (Lin, 2025), which enables better results. Second, the study focuses on the interaction between the Corpus Linguist model and corpus data. The aim is to evaluate both the potential and current limitations of combining GenAI and corpus data in

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 80–86
November 20-22, 2025

terms of corpus searching and the generation of corpus-based exercises.

2 Data

For the purpose of this study, the ChatGPT (OpenAI, 2025) web interface and the GPT model Corpus Linguist (Milička & Machálek, 2024) were used. The Corpus Linguist model enables communication with the resources of the Czech National Corpus (CNC) project (Křen et al., 2016), which is an umbrella term for many corpora, not only those delivering data in the Czech language. The GPT model communicates through an API with written corpora (SYN2020, SYN v12), spoken corpora (ORTOFON), and parallel corpora (InterCorp) covering most European languages such as Croatian, Czech, Danish, Dutch, German, English, Finnish, French, Greek, Hungarian, Italian, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish. It also allows communication with the corpus of Czech contemporary poetry, the British National Corpus, and the Old Bailey Corpus. Of course, the CNC offers many more corpora available via its website (www.korpus.cz), but only the aforementioned ones are accessible through the API.

3 Workflow with Chat-GPT

In this study, the automatic generation of corpus-based exercises makes use of Czech corpora to evaluate exercises intended for foreign learners of Czech. However, all communication with the GPT model was carried out in English. The first step was to determine what kind of corpus-based exercises the model is able to generate. Next, the workflow with the model focused on a single exercise type in order to evaluate the interaction between the language model and corpus data.

Based on previous studies (Lin, 2025), it is known that the more detailed prompt is, the more plausible the results are. Therefore, communication with the Corpus Linguist model started with a detailed question: *Could you list the types of language exercises for foreigners that you are able to create based on corpus data?* In return, six categories of exercises were obtained:

1. Vocabulary & Collocations

- <u>Fill-in-the-blank with collocation</u>

  example: "He made a ___ decision" (from frequent collocates of decision).

- <u>Match words with their typical partners</u>

  example: *strong tea* vs. **powerful tea*.

- <u>Frequency-based vocabulary lists</u>

  most common words in a domain, genre, or register.

2. Grammar Practice

- <u>Conjugation & declension in context</u> sentences from corpus with a verb missing (*oni* ___ *domů* → *šli*, 'they went home').

- <u>Case exercises (Czech)</u>

  fill the missing case ending from authentic examples.

- <u>Word order practice jumbled authentic</u> sentences to be reordered correctly.

3. Reading & Comprehension

- <u>Gap-fill reading texts</u>

  real corpus sentences with systematic gaps (e.g., prepositions, verbs).

- <u>Multiple-choice from concordances</u> choose which option fits the authentic sentence.

- <u>Register awareness</u>

  compare informal spoken excerpts (ORTOFON) with formal written ones (SYN).

4. Error Detection & Correction

- <u>Spot the odd one out</u>

  one untypical collocate among three authentic ones.

- <u>Learner error simulations</u>

  based on what corpus shows is not used, learners choose the correct form.

5. Translation & Contrastive Tasks (using InterCorp)

- <u>Parallel text exercises</u>

  learners match Czech sentences with their English (or other language) equivalents.

- False friends & tricky equivalents

  highlight differences between corpora in two languages.

6. Communicative / Production

- Prompts with authentic phrases

  learners continue a sentence taken from corpus.

- Roleplay triggers

  situational expressions mined from spoken corpora ("Můžu si přisednout?", 'Can I sit down?').

The model indicated that all these exercises can be data-driven and offered to prepare a sample exercise. To ensure that all possible exercises were listed, the model was asked: *Did you listed all possible exercises?* In response, the model honestly answered that it was a representative set of exercises and that more possibilities are available depending on teaching goals, learners' level, and other factors. It then provided additional exercise types as follows:

1. Lexical / Phraseological

- Idioms & proverbs

  find authentic variants and ask learners to complete or explain them.

- Synonym choice

  pick the most frequent synonym in a given context.

2. Morphosyntax

- Transformation tasks

  rewrite corpus sentences (active → passive, direct → reported speech).

- Agreement practice

  highlight mismatches and ask learners to fix them (subject–verb, adjective–noun).

- Tense/aspect recognition (Czech perfective vs. imperfective verbs).

3. Discourse / Pragmatics

- Dialogue completion

  fill missing turns in authentic spoken dialogues.

- Register/style awareness

  classify sentences as formal/informal.

- Connector use

  identify or insert discourse markers (*totiž*, *nicméně* 'however').

4. Writing Support

- Sentence starters

  give frequent corpus-based openings, learners continue.

- Chunk recycling

  learners create new sentences with frequent multiword units.

- Error-avoidance training

  contrast common learner mistakes with authentic usage.

5. Assessment & Awareness

- Acceptability judgments

  choose the corpus-attested variant.

- Frequency awareness

  rank synonyms according to corpus frequency.

- Guess the meaning from context

  show concordances with a new word hidden.

Two exercise types were chosen for the further evaluation of the interaction between the language model and corpus data. The first was an exercise on error detection, where a learner needs to identify one untypical collocate among three authentic ones, and the second focused on false friends. The model was also asked to provide information about corpus data use.

Table 1 and Table 3 present a generative result of GenAI, divided into three steps: 1) *Exercise*, 2) *Task*, and 3) *Answer*. In both cases, a target word was not specified, so the model selected a lemma on its own.

In the first exercise, the chosen lexeme was *rozhodnutí* 'decision'. The GPT model explained the individual steps, stating that the top authentic collocates were provided based on the SYN2020 corpus and that one example was added as the "odd one out", an unattested collocate. However, a discrepancy can be observed between the collocates in the Exercise and Task parts. The collocation *učinit rozhodnutí* ('make a decision') does not appear in the Task part and is replaced by *těžké rozhodnutí* (‘hard decision’).

<table>
<tr><td>Prompt</td><td>Could you generate an exercise type "Spot the odd one out"? Please inform me, what kind of corpus data you used.</td></tr>
<tr><td colspan="2">Exercise: Spot the odd one out (Czech collocations)</td></tr>
<tr><td colspan="2">Target word: lemma rozhodnutí (“decision”).
• From the SYN2020 written corpus, I can extract its most frequent collocates.
• Top authentic collocates include: učinit rozhodnutí (“make a decision”), důležité rozhodnutí (“important decision”), správné rozhodnutí (“right decision”).
• I will then add one implausible or unattested collocate as the “odd one out”.</td></tr>
<tr><td colspan="2">Task</td></tr>
<tr><td colspan="2">Which of the following combinations is NOT typical in Czech? Circle the odd one out.
1. důležité rozhodnutí
2. správné rozhodnutí
3. červené rozhodnutí
4. těžké rozhodnutí</td></tr>
<tr><td colspan="2">Answer</td></tr>
<tr><td colspan="2">3. červené rozhodnutí (“red decision”)
• It is not attested in corpus collocations and does not make semantic sense.</td></tr>
</table>

Table 1: Generative result of corpus exercise "Spot the odd one out".

To verify that the model was indeed communicating with the corpus, the most frequent collocates of the lexeme *rozhodnutí* ‘decision’ were extracted through the corpus portal in the same way as described by the model, using the logDice collocation measure, a span window of ±5 words, and a minimum frequency of 3 within the collocation. *Správné* (‘right’) was the only collocate provided by GenAI in the exercise that also appeared in the top 10 collocates (Table 2) based on the corpus search. It ranked 8th, while the next two appeared at ranks 43 (*důležité* ‘important’) and 62 (*těžké* ‘hard’).

The model stated that it “chose 3 real frequent collocates and added 1 fabricated/unattested one as the distractor.” However, not all collocates were among the most frequent. Therefore, the model was asked to deliver the collocation profile of the lemma *rozhodnutí* from the written Czech corpus SYN2020 (Křen et al., 2020). The results were identical to those of the corpus search, which confirmed the communication between the GPT model and the corpus portal.

Rank	Lemma	Traslation (POS)	Fq	logDice
1.	soud	court (N)	940	9.72
2.	územní	territorial (A)	305	9.11
3.	učinit	to make (V)	354	9.08
4.	soudní	judicial (A)	267	8.73
5.	padnout	to fall (V)	303	8.66
6.	dospět	to come (V)	174	8.18
7.	pravomocný	authoritative (A)	135	8.17
8.	správný	right (A)	285	8.08
9.	činit	to make (V)	190	8.03
10.	konečný	final (A)	203	8.01

Table 2: Top 10 collocate of lexeme *rozhodnutí* in SYN2020.

The most top collocates of lemma *rozhodnutí* are mostly related to an institutional context, which might be the reason why the model did not use this vocabulary, as it is not always pedagogically useful for lower proficiency levels. The model also argued that semantically clear collocations are needed for pedagogical reasons. This reasoning is valid, and the procedure for the automatic generation of corpus-based exercises provides not only authentic examples for learners but also the most appropriate ones from a pedagogical perspective. The longer the communication with the GPT model is conducted, the better the results obtained, and the interaction with corpus data is well explained.

In the second exercise (Table 3), the model chose the interesting Czech lexeme *eventuálně*, that is a false friend with English lexeme *eventually*. It explained that the parallel corpus InterCorp v13ud (Čermák & Rosen, 2012) was used to extract concordance lines for *eventuálně* in Czech texts aligned with English translations. The model also assumed that corpus frequency results identified *possibly* and *alternatively* as the major equivalents. However, a direct consultation of the corpus revealed discrepancies. Following the same steps described by the model, 156 hits of *eventuálně* were found in the Czech-English part of InterCorp v13ud (Klégr et al., 2021). Manual analysis of the context confirmed that the most frequent equivalent was *possibly*, with 33 hits, however, it was followed by *eventually* with 28 (in the subtitle collection 14) hits (Figure 1). *Alternatively* appeared only 7 times. The other results were mostly translated by omission or by other equivalents not mentioned by the model. Further prompting revealed that the model is not able to consult corpus data when it comes to two parallel corpora aligned directly. The current API settings do not allow this function. In fact, the model provided misleading information, claiming that the corpus data delivered the most frequent translations. This type of hallucination is particularly problematic, as the Corpus Linguist model was expected to cooperate with corpus data. Nonetheless, it highlights the unreliability of some outputs, which users should be aware of.

Prompt	*Could you generate an exercise type "False friends & tricky equivalents"? Please inform me, what kind of corpus data you used.*
Exercise: Exercise: False Friends (Czech–English)	
Target word: Czech *eventuálně* • Many learners think it means English *eventually*. • But in corpus data, *eventuálně* is most often translated as *possibly / alternatively*.	
Task	
Fill in the blanks with the correct English equivalent of Czech *eventuálně*. Choose between **eventually** and **possibly**. *Pokud se nám to nepodaří, můžeme ___ zvolit jinou metodu.* → If we don't succeed, we can ___ choose another method. *Po dlouhých diskusích ___ dospěli k rozhodnutí.* → After long discussions, they ___ reached a decision. *Tyto části mohou být ___ nahrazeny novými.* → These parts can be ___ replaced by new ones.	
Answer	
1. possibly 2. eventually 3. possibly	

Table 3: Generative result of corpus exercise "False friends & tricky equivalents".

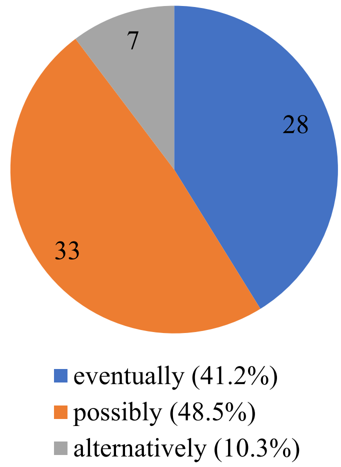


Figure 1: Distribution of translation equivalents of lexeme *eventuálně* in InterCorp 13ud.

The exercise itself asks learners to fill in the gaps in English sentences, not in Czech ones. Therefore, it resembles practice for Czech native speakers studying English rather than exercises for foreign learners of Czech. Moreover, the second Czech sentence should not contain the lexeme *eventuálně* but *nakonec* 'in the end'. This exercise requires improvement in this regard. Further prompting produced a new version of the exercise in which learners were asked to choose between two English sentences, e.g. **The trip can eventually be extended* and *The trip can possibly be extended*, to identify the correct equivalent of the given Czech sentence *Cestu lze eventuálně prodloužit*.

4 Discussion

This study, based on two examples of corpus-based exercises, explored the possible communication of a GPT model (Corpus Linguist) with corpus data (CNC). The results revealed that the model successfully extracted data from corpora in some cases, however, in other cases, it claimed to use corpus data that were not actually consulted. Comparison with previous a study (Zasina, 2025) also indicates that the Corpus Linguists model tends to choose similar target lexemes (*rozhodnutí* 'decision') for different exercises when a target lexeme is not specified in the prompt. The model appears to be pre-set to choose target words for exercise generation, which represents a certain limitation of the tool. Therefore, for the automatic generation of corpus-based exercises, it is essential to deliver input data concerning the most problematic areas for the learners for whom the exercises are intended. This precaution has the potential to improve the generated results and better target learner-specific needs. It seems that the model is not yet able to truly consider learners' real needs, and this remains the task of materials developers.

The two examples also demonstrated that it is crucial to define detailed prompts and verify the answers through subsequent prompting. Longer interactions make it possible to identify the processes undertaken by the GPT model. This is an important consideration in prompt crafting, to avoid succumbing to the illusion that every answer is correct. It should be remembered that GenAI can only produce strings of characters that form words and sentences; it is not capable of independent thought.

An interesting concept of GenAI as a role player (see Shanahan et al., 2023) can be applied in this context. When prompting a GPT model, one may feel that the automatically generated sentences are human-like. However, GenAI's ability to "act" convincingly stems from its vast training data (Shanahan et al., 2023, p. 496), and it strives to deliver coherent responses. Some scholars (Milička, 2024, p. 16) have emphasised that the model should not be anthropomorphised. Anthropomorphising GPT models can be a trap that dulls vigilance. Thus, it is important to critically assess GenAI's outputs, especially when combining them with corpus data.

Some attempts to integrate GenAI with corpus data in language learning have been undertaken (Cheung & Crosthwaite, 2025; Crosthwaite & Anthony, 2025, p. 6; Zasina, 2025) and have produced promising results. However, this study highlighted its limitations regarding communication with corpus data via the ChatGPT interface, which tends to provide an answer under any circumstances. Users should therefore be cautious and prepare more elaborated prompts that explicitly query the source data.

Even though this evaluation is limited to two examples of corpus-based exercises, it provides insight into how GenAI interacts with CNC sources. It offers guidance for future developments in prompt crafting and for evaluating the reliability of generative results. Furthermore, it underscores that users should approach GenAI critically. Further investigation may lead to improvements that could mitigate these issues.

5 Conclusion

This study focused on the interaction between language models and corpus data in terms of DDL practice. It shows that there are many possibilities for language learners to combine GenAI and corpus data. GenAI can effectively use corpus data within GPT models to produce corpus-based exercises. However, it is important to emphasise that prompts should contain high-quality input information and be as precise as possible in order to obtain appropriate results. It is also crucial to critically evaluate the automatically generated outputs, as they can be hallucinated by GenAI. In the future, further evaluations of this kind will be necessary to determine whether AI sufficiently cooperates with corpus data or merely creates the illusion of real corpus results.

Acknowledgments

This work was supported by the project "Human-centred AI for a Sustainable and Adaptive Society" (reg. no.: CZ.02.01.01/00/23_025/0008691), co-funded by the European Union, and the project "Multilingual Lens: Investigating Large Text Corpora from Different Methodological Perspectives" (UNCE/24/SSH/009).

References

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, *17*(3), 411–427.

Cheung, L., & Crosthwaite, P. (2025). *CorpusChat*: Integrating corpus linguistics and generative AI for academic writing development. *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2025.2506480

Crosthwaite, P., & Anthony, L. (2025). Tools for Data-Driven Learning. In *The Palgrave Encyclopedia of Computer-Assisted Language Learning* (pp. 1–9). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-51447-0_74-1

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *Classroom Concordancing: ELR Journal*, *4*, 1–16.

Klégr, A., Kubánek, M., Malá, M., Rohrauer, L., Šaldová, P., Šebestová, D., Vavřín, M. & Rosen A. (2021). *InterCorp – English, Release 13ud of 22 December 2021*. Institute of the Czech National Corpus, Charles University. www.korpus.cz

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P. & Zasina, A. J. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2522–2528). Portorož: ELRA.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., & Škrabal, M. (2020). *SYN2020: Representative corpus of written Czech*. Institute of the Czech National Corpus, Faculty of Arts, Charles University. www.korpus.cz

Lin, P. (2025). AI Chatbots and Data-Driven Learning. In L. McCallum & D. Tafazoli (Eds.), *Encyclopedia of Computer-Assisted Language Learning* (pp. 1–8). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-51447-0_67-1

Milička, J. (2024). *Theoretical and Methodological Framework for Studying Texts Produced by Large Language Models* (arXiv:2408.16740). arXiv. https://doi.org/10.48550/arXiv.2408.16740

Milička, J., & Machálek, T. (2024). *Corpus Linguist* (Version January 17, 2025) [Computer software]. https://chatgpt.com/g/g-pFqRCNeHu-corpus-linguist

OpenAI. (2025). *Chat-GPT* (Version September 9, 2025) [Computer software]. https://chatgpt.com/

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Vyatkina, N. (2020). Corpora as open educational resources for language teaching. *Foreign Language Annals*, *53*(2), 359–370. https://doi.org/10.1111/flan.12464

Zasina, A. J. (2022). Designing a Corpus Workbook for Students of Czech as a Foreign Language. *Studie z Aplikované Lingvistiky - Studies in Applied Linguistics*, *13*(2), 125–132.

Zasina, A. J. (2025). Typologie korpusových cvičení a jejich automatické generování pomocí AI [Typology of corpus-based exercises and their automatic generation using AI]. In M. Škrabal, B. Štěpánková & H. Skoumalová (Eds.), *Korpus třicetiletý* (pp. 127−145). Praha: Nakladatelství Lidové noviny.

Diversity is the Key: Enhancing LLM-based Post-processing for Automated Audio Captioning

Seyed Ali Farokh, Mohammad Mehdi Homayounpour, Ahmad Nickabadi
Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran
{alifarokh, homayoun, nickabadi}@aut.ac.ir

Abstract

Automated Audio Captioning (AAC) is a multimodal task aimed at generating natural language descriptions of audio content. Previous studies have shown that LLMs can improve AAC performance by summarizing audio events based on a list of candidate captions, which are selected by an external reranker from those generated using Nucleus Sampling. However, the reranking process often selects overly similar captions, disregarding the original diversity of the sampled captions. In this work, we show that this diversity reflects the AAC model's level of certainty and propose a lightweight candidate selection approach that preserves the initial diversity of the generated captions. This, in turn, enables an LLM to summarize the captions while considering the AAC model's certainty in a few-shot setting. Experimental results demonstrate that our method outperforms previous post-processing techniques while being significantly faster.

Keywords: Automated Audio Captioning, Large Language Models, In-context Learning, Post-processing

1 Introduction

Automated Audio Captioning (AAC) is a multimodal task that aims to generate natural language descriptions of the content within audio samples. Unlike Automatic Speech Recognition (ASR), which focuses on transcribing spoken language, AAC primarily targets environmental and non-speech sounds, providing meaningful descriptions of auditory scenes and events.

One of the primary challenges in AAC lies in the inherent ambiguity of audio signals. Unlike image captioning, where objects can be described through concrete attributes such as shape, color, size, and spatial relationships, describing audio clips is significantly more complex (Wu et al., 2019). Acoustic events often exhibit overlapping or similar sound characteristics, leading to varied perceptions among individuals, as shown in Figure 1 (Zhang et al., 2023; Drossos et al., 2020). Consequently, widely used audio captioning datasets, such as Clotho (Drossos et al., 2020), provide multiple ground-truth captions from different annotators for each audio sample, and models are typically trained on one-to-many audio-caption pairs, where each audio clip is randomly paired with a single ground-truth caption in each iteration (Zhang et al., 2023). This can introduce uncertainty in the learned representations and potentially result in performance degradation.

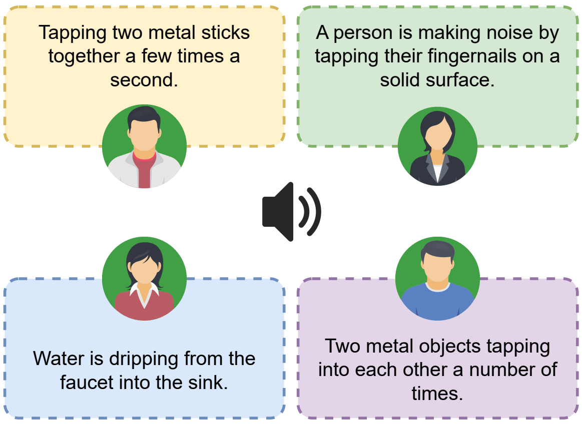


Figure 1: Diverse and occasionally conflicting perceptions of a single audio sample due to the inherent ambiguity of audio signals. The captions correspond to one training sample from the Clotho dataset (`dual metal.wav`).

Thanks to the annual DCASE challenges[1] and the release of open-source audio captioning datasets such as Clotho (Drossos et al., 2020) and AudioCaps (Kim et al., 2019), advancements in AAC research have gained momentum in recent

[1] IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, available at https://dcase.community

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 87–94
November 20-22, 2025

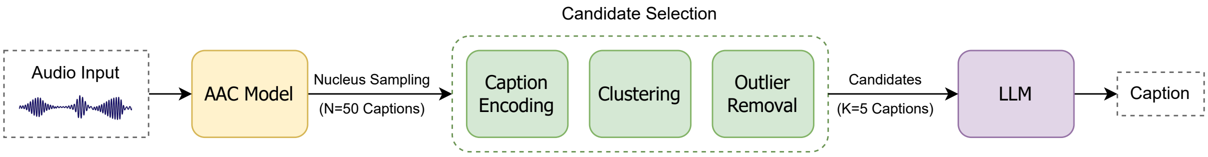


Figure 2: Overview of our proposed method. First, $N = 50$ captions are generated for a given audio input using Nucleus Sampling. Next, in the candidate selection stage, $K = 5$ captions are chosen to preserve the diversity of the generated captions. Finally, these selected captions are processed by an LLM to further enhance diversity and produce the final caption.

years. Most state-of-the-art AAC models employ an encoder-decoder architecture (Xu et al., 2022; Ye et al., 2022; Narisetty et al., 2021; Wu et al., 2024), where the encoder is typically a pre-trained audio encoder, such as PANN (Kong et al., 2020) or BEATs (Chen et al., 2023), that extracts audio features from the input signal. These features are then passed to an autoregressive text decoder, such as BART (Lewis et al., 2020) or GPT-2 (Radford et al., 2019), which generates the corresponding caption. The decoders normally generate sequences using greedy decoding and beam decoding.

In addition to these conventional decoding methods, recent research has demonstrated that a hybrid sampling and reranking strategy, which leverages external pre-trained models for reranking, can improve the outputs of AAC models by exploring a broader search space than beam search (Wu et al., 2024; Jung et al., 2024). Furthermore, inspired by the success of Large Language Models (LLMs) in a zero-shot setting across a variety of tasks and their ability to generate human-like text (Radford et al., 2019), recent studies in AAC have incorporated zero-shot LLM-based caption summarization (Jung et al., 2024) and error correction (Liu et al., 2024) as post-processing steps, demonstrating the effectiveness of these techniques in refining the generated captions.

In this work, we hypothesize that the diversity of sampled captions reflects the AAC model's level of certainty regarding a given input. We demonstrate that reranking is not the most effective approach for candidate caption selection, as the resulting captions lack sufficient diversity to both capture the model's uncertainty and serve as input for LLM-based summarization. To address this limitation, we propose a method that preserves the original diversity of the sampled captions and employs an LLM in a few-shot setting to generate a final caption while considering the AAC model's uncertainty. Experimental results show that the proposed method outperforms previous post-processing techniques while being significantly simpler and faster. Our contributions can be summarized as follows: (1) we propose a lightweight candidate caption selection method as an alternative to the hybrid sampling and reranking strategy, (2) we enhance AAC performance through an LLM-based post-processing approach that leverages in-context learning and accounts for the AAC model's certainty, and (3) we introduce a simple technique to identify high-quality captions generated by AAC models, enabling selective LLM-based refinement that improves performance while minimizing unnecessary processing overhead. The source code, experiments, and results from this work will be publicly released upon publication.

2 Related Work

2.1 Sampling and Reranking

Nucleus Sampling (Holtzman et al., 2020), also known as Top-p Sampling, is a stochastic decoding method commonly used in modern LLMs to generate more diverse text compared to conventional beam decoding (Shi et al., 2024). Inspired by this, Wu et al. (2024), the winners of the DCASE 2023 AAC challenge, observed that approximately one-third of the captions generated using Nucleus Sampling achieve higher scores on AAC metrics compared to those produced with beam decoding. To leverage this advantage, they proposed a sampling and reranking strategy that first generates a list of N captions (50 in their original work) using Nucleus Sampling, followed by a hybrid reranking method to select the most suitable caption from the list by computing two reranking scores. The *decoder score* is obtained by feeding the input waveform into the encoder and the generated captions into the decoder to directly compute the caption log-likelihood on the decoder outputs. The *encoder score* is calculated as the cosine similarity between the audio embedding, obtained by feeding the input

Audio File: Shanghai Traffic Near Peoples Square.wav		
	CLAP Reranking	**Our Proposed Method**
Candidate Selection	Several cars and trucks are driving by on a busy street. Several cars driving by on a busy street. A busy street with cars driving by on a nearby road. A busy street with vehicles passing by. A large number of cars pass by on a nearby road.	A car drives by and then stops. Several cars are driving by on a busy road. Cars are driving by on the street and a woman is speaking. A person is walking down the street with cars driving by. A car drives by and people are talking.
LLM Output (FENSE)	Several cars and trucks are driving by on a busy street. (53.67%)	Cars are driving by on a busy road while a person is walking and people are talking. (70.35%)
Explanation	CLAP Reranking ranks the sampled captions based on their embedding similarity to a fixed-dimensional audio embedding, capturing only the aspects of the audio emphasized by the CLAP audio encoder. This process may overlook some events detected by the AAC model, such as the sound of people's voices in this example. In contrast, our proposed method selects the most salient concept clusters, better reflecting the full range of events recognized by the AAC model.	
Audio File: Fountain Trompenburg 090928.wav		
	CLAP Reranking	**Our Proposed Method**
Candidate Selection	A stream is flowing over rocks as people chatter and walk. Water is flowing in a creek as people talk and walk. Water is flowing as people talk and walk by. Water is flowing as people talk and walk through a stream. A stream of water flows while people talk and walk.	Water is flowing down a stream as people talk in the background.
LLM Output (FENSE)	Water is flowing in a stream as people talk and walk by. (45.77%)	Water is flowing down a stream as people talk in the background. (53.31%)
Explanation	The low diversity among the sampled captions in this example indicates that the AAC model was highly confident about the events in the audio. This is further supported by the fact that our method identified only a single salient cluster. As a result, we skip LLM inference and directly use the centroid of this cluster as the final caption. This not only reduces computational overhead but may also improve evaluation scores, as the AAC model is trained to align with the target caption distribution, whereas the LLM, operating in a few-shot setting, is less familiar with the characteristics of AAC-generated captions.	

Table 1: Illustration of how different candidate selection methods affect the LLM's output.

waveform into the encoder, and the caption embedding, derived by feeding the generated caption into a pre-trained text encoder, i.e., INSTRUCTOR (Su et al., 2023). Finally, the generated captions are reranked using a weighted sum of the *decoder* (0.3) and *encoder* (0.7) scores, with the top-ranked caption selected as the system output.

However, our experiments revealed that the *decoder score* has a negligible effect and can be safely omitted without significantly impacting performance. Specifically, the system achieves FENSE scores of 52.13 and 50.17 when using only the *encoder* or *decoder score* for reranking, respectively, while the fused scores yield a performance of 52.28. This suggests that the success of the proposed reranking method relies heavily on the additional supervision signal provided by INSTRUCTOR during training, which prevents it from being applied to other pre-trained AAC models.

In DCASE 2024, Jung et al. (2024) introduced a model-independent reranking approach based on CLAP (Wu et al., 2023), a multimodal audio and text encoder that uses contrastive learning techniques to jointly embed these two modalities. Their approach is similar to the previous sampling and reranking method, with the key difference that they encode both the generated captions and the input audio using CLAP. Additionally, beyond utilizing CLAP for reranking, they proposed incorporating it as an additional filtering stage prior to the previously described hybrid reranking method. This filtering step removes half of the generated captions that are not sufficiently aligned with the audio embedding.

2.2 LLM-based Summarization

Given that LLMs have been proven effective across a range of zero-shot tasks, Jung et al. (2024) adopt an LLM-based caption summarization method. In this approach, a sampling and reranking strategy is first used to rank a set of sampled captions. Next, rather than selecting the top-ranked caption, the top-K captions are fed into an LLM with a zero-shot caption summarization prompt to generate the final caption. This method aims to enrich the final caption by combining key phrases that may be scat-

Prompt Template
You are provided with several candidate captions generated by an Automated Audio Captioning system for a specific audio file. These captions may contain repetitions, inaccuracies, or illogical details. Each caption may describe one or more main events. Identify the most frequent and relevant events from all the captions, and generate a single caption, logically describing the most probable events present in the original audio. Ensure the caption is free of punctuation marks, including commas. **Captions:** A car is driving down a road with the window open. The rain is falling as a car passes by. Water is flowing as a car passes by. The rain is falling and the wind is blowing. **Generated Caption:** A car is passing by while the rain is falling and the wind is blowing. **Captions:** Cars are passing by on a busy road. Cars drive by on a busy highway while a wind blows. Cars drive by on a wet road. A car is driving down the road and then the car drives by. **Generated Caption:** Cars are driving down a busy and wet road while the wind blows. *[more demonstrations]* **Captions:** *[selected candidates]*

Table 2: Few-shot prompt template.

tered across different sampled captions, while also leveraging the LLM's ability to generate grammatically accurate sentences. However, in our experiments, we observe that the reranking stage considerably diminishes the diversity of the selected captions, often resulting in many identical captions, thereby reducing the effectiveness of LLM-based summarization.

2.3 LLM-based Error Correction

In their recent work, Liu et al. (2024) used an LLM as a post-corrector to address potential grammatical errors and repetitions in the captions generated by their AAC model, operating in a one-shot setting. In this approach, only a single caption sample from the AAC model is provided to the LLM for error correction.

3 Methodology

A major challenge in AAC arises from the inherent ambiguity of audio signals. Due to the overlapping and similar sound characteristics of many acoustic events, individuals may perceive the same audio differently, sometimes even with conflicting interpretations (Figure 1). To address this variability, popular audio captioning datasets, such as Clotho, provide multiple ground-truth captions from various annotators for each audio sample (Drossos et al., 2020). During training, models are exposed to one-to-many audio-caption mappings, with each audio clip paired with a randomly selected ground-truth caption in each epoch. This randomness can introduce uncertainty into the learned representations and degrade model performance (Zhang et al., 2023).

To examine how this uncertainty affects the output of AAC models, we randomly selected 50 audio samples from the Clotho dataset and generated 50 captions per audio sample with Nucleus Sampling using two pre-trained AAC models. A careful manual analysis of the generated captions revealed that the AAC model's confidence in the acoustic events of a given input audio is strongly reflected in the diversity of the sampled captions. Specifically, when the AAC model is confident about the audio content, nearly all sampled captions describe the same events, differing only in word choice and ordering. Conversely, when the input audio is ambiguous or challenging, the sampled captions display greater diversity, describing a range of possible events.

Thus, we hypothesize that the diversity of sampled captions can serve as an indicator of an AAC model's confidence level. Based on this hypothesis, we propose a post-processing method for AAC models with the following steps (Figure 2): First, we generate N captions for each input audio using Nucleus Sampling and encode them with a lightweight sentence encoder. Next, the encoded captions are clustered into K groups to identify the primary event clusters. The K cluster centroids, representing the primary possible events, are then fed into an LLM along with a few demonstrations to generate the final caption. When the selected captions describe similar events, the LLM is expected to produce a consistent caption with its inputs. However, when the diversity among the selected captions is high, the LLM should incorporate different possible events, resulting in a more diverse and comprehensive output. The following subsections provide a detailed explanation of each step, and Table 1 presents two illustrative examples.

3.1 Sampling and Candidate Selection

For each given input audio, we use Nucleus Sampling to generate a set of N diverse captions. We

AAC Model	Decoding & Post-Processing	FENSE (%)
CoNeTTE (Labbé et al., 2024)	Beam Decoding (width=5)	51.96
CoNeTTE	Beam Decoding (width=5) + LLM-based Error Correction (Liu et al., 2024)	51.60
CoNeTTE	Sampling + CLAP Reranking (Jung et al., 2024)	49.86
CoNeTTE	Sampling + CLAP Reranking + LLM-based Summarization (Jung et al., 2024)	53.32
CoNeTTE	Sampling + Ours	**53.76**
BEATs-Conformer-BART (Wu et al., 2024)	Beam Decoding (width=5)	50.35
BEATs-Conformer-BART	Beam Decoding (width=5) + LLM-based Error Correction	50.15
BEATs-Conformer-BART	Sampling + Hybrid Reranking (Wu et al., 2024)	52.28
BEATs-Conformer-BART	Sampling + Hybrid Reranking + LLM-based Summarization	52.63
BEATs-Conformer-BART	Sampling + CLAP Reranking	51.49
BEATs-Conformer-BART	Sampling + CLAP Reranking + LLM-based Summarization	52.71
BEATs-Conformer-BART	Sampling + CLAP Filtering + Hybrid Reranking	52.75
BEATs-Conformer-BART	Sampling + CLAP Filtering + Hybrid Reranking + LLM-based Summarization	52.89
BEATs-Conformer-BART	Sampling + Ours	**53.49**

Table 3: Results on the evaluation subset of Clotho.

then select a set of $K = 5$ candidate captions that preserve the original diversity of events present in the generated captions (the first example in Table 1). To achieve this, we use a lightweight pre-trained off-the-shelf text encoder, Sentence-BERT (Reimers and Gurevych, 2019), to encode the captions into vector embeddings, and then apply Agglomerative clustering with complete link to group them into K clusters. For each cluster, we compute the center point by averaging the embeddings of the captions within the cluster, and select the caption with the closest embedding to this center as the cluster representative. In this work, cosine similarity was used consistently across all embedding-based steps.

Moreover, to prevent the selection of too infrequent events that could mislead the LLM, we incorporate an outlier removal step during this phase, removing clusters with fewer than $R = 5$ embeddings. Additionally, when the majority of embeddings fall into a single cluster (at least $C = 72\%$ of the embeddings), indicating high confidence from the AAC model, we bypass the LLM phase and directly use the cluster representative as the system output (the second example in Table 1). This approach not only reduces the computational overhead of LLM inference but also enhances performance, as the AAC model is specifically trained to generate captions and is more adept at producing outputs that align with the target distribution. This simple yet effective step is also extendable to other LLM-based post-processing methods.

3.2 Few-shot Caption Diversity Enhancement

The selected captions are then processed by an LLM using a few-shot prompt to generate the final caption. When there is high diversity among the input candidate captions, the LLM is anticipated to generate a more diverse caption. Conversely, when the diversity is low, the LLM is expected to produce a caption that closely matches the inputs. Table 2 contains the prompt template used for this task. Since the primary goal of this study is to evaluate the impact of diversity-enhanced candidate selection, we did not focus on optimizing the number or content of demonstrations used in the LLM prompt. Instead, a fixed set of five manually crafted demonstrations was used across all inputs. This choice was supported by preliminary experiments, which indicated that four to six demonstrations are generally sufficient for reasonable LLM performance, depending on the model. Given that manually creating this small number of examples is straightforward, we leave the exploration of automatic demonstration optimization for future work. The complete list of demonstrations can be found in the accompanying source code.

4 Experimental Setup

4.1 Models

Our proposed post-processing method is independent of the AAC model. Thus, we conduct our experiments using two open-source models: CoNeTTE[2] (Labbé et al., 2024) and BEATs-Conformer-BART[3] (Wu et al., 2024). Additionally, GPT-4o-mini is used as the LLM in our experiments, accessed through the OpenAI API.

[2]https://github.com/Labbeti/conette-audio-captioning

[3]https://github.com/slSeanWU/beats-conformer-bart-audio-captioner

4.2 Hyperparameters

During the sampling phase of all methods, Nucleus Sampling was performed with a temperature of 0.5 and a top-p value of 0.95. Greedy decoding was used for all LLM-based stages. The max tokens parameter was set to 50 for both Nucleus Sampling and LLM generations.

The parameters $K = 5$ and $R = 5$ were selected based on intuition and preliminary experiments. We observed that moderate changes to these values do not significantly affect the results, and the chosen values offer a good balance that works well across a wide range of AAC models and LLMs. In contrast, $C = 0.72$ was determined through grid search on Clotho's validation subset.

4.3 Dataset

We conduct our experiments using the Clotho v2.1 dataset (Drossos et al., 2020), which served as the standard benchmark in previous DCASE scientific challenges. The dataset consists of four subsets. The *development* and *validation* subsets are intended solely for optimizing AAC models, while the *evaluation* subset is used for assessing and comparing results. The *testing* subset is reserved exclusively for scientific challenges, such as the DCASE challenge. To conform with this standard, we use only the *evaluation* subset of the dataset to compare and report our results.

4.4 Evaluation Metrics

We adopt the FENSE metric (Zhou et al., 2022), the standard evaluation metric of the DCASE 2024 challenge, as our evaluation metric. Prior to FENSE, AAC evaluation metrics were borrowed from machine translation and image captioning and focused on the surface form of the words (Labbé et al., 2024). FENSE, on the other hand, leverages pre-trained models to capture sentence meanings. It also penalizes grammatically incorrect or incoherent sentences.

5 Results and Discussion

The experimental results (Table 3) demonstrate the effectiveness of our proposed method compared to other post-processing approaches when applied to the outputs of two open-source AAC models. These findings underscore the importance of preserving the diversity of sampled candidates, particularly for LLM-based post-processing methods.

Method	FENSE (%)
Random Selection (K=5)	52.46
Random Selection (K=20)	53.05
All Candidates (K=50)	53.07
Clustering (K=5)	52.21
+ Outlier Removal	53.31
+ Skipping LLM Usage	**53.49**

Table 4: Ablation study of the candidate selection stages.

Additionally, Table 1 presents two concrete examples that illustrate how our method works in practice and provide intuition behind its effectiveness. In these examples, the same prompt template was used across different candidate selection methods to enable a fair comparison, ensuring that the observed improvements can be attributed solely to the proposed candidate selection strategy rather than differences in prompt design compared to prior studies.

5.1 Ablation Study

We conduct a comprehensive ablation study on the candidate selection phase, beginning with a random candidate selection method and gradually incorporating the proposed components. Table 4 shows that the clustering phase is significantly affected by outliers, leading to performance that falls behind random candidate selection. However, removing the outliers results in a notable improvement, emphasizing the importance of this step. Additionally, while including more samples in the prompt, up to selecting all generated captions, can slightly improve performance, it still lags behind the proposed clustering method. This is likely due to the large volume of redundant information the LLM must process, as well as the presence of outliers that represent highly unlikely events in the inputs. These findings underscore the importance of targeted candidate selection. Finally, skipping LLM inference when a single cluster contains more than C captions leads to additional performance gains. In this specific scenario, although the overall improvement across the entire Clotho evaluation set may appear modest, the FENSE score increases from 55.15 to 56.11 for the 116 samples where this condition applies (approximately 11% of the subset). This demonstrates that the method effectively identifies cases where the AAC model is confident

Stage	Time (ms)
Beam Decoding (width=5)	487
Nucleus Sampling (N=50)	991
Hybrid Reranking (Wu et al., 2024)	739
CLAP Reranking (Jung et al., 2024)	340
CLAP Filtering + Hybrid Reranking	833
Candidate Selection (Ours)	15
LLM Inference (GPT-4o-mini)	779

Table 5: Average processing time per sample (in milliseconds) for various decoding and post-processing methods.

and avoids unnecessary LLM processing.

5.2 Runtime Analysis

As depicted in Table 5, our proposed candidate selection stage is considerably faster than previous reranking strategies. The processing times were calculated by running the methods on the entire evaluation subset of Clotho v2.1 using a machine with a single Nvidia RTX 3090 GPU. The LLM inference time, which includes the HTTP request and response times as well, was measured on Google Colaboratory servers. Since this time was consistently similar across different inputs, with negligible variations, only the average time is reported. During each stage, parallelism was disabled, and all samples were processed sequentially. The AAC model used throughout all stages was BEATs-Conformer-BART.

6 Conclusion and Future Work

In this work, we explored various post-processing methods for automated audio captioning and proposed a novel LLM-based method for enhancing caption diversity. The proposed approach leverages in-context learning to consider the certainty of the AAC model, reflected in the diversity of its generated captions. Despite being considerably faster, our method demonstrates performance improvements over previous post-processing techniques, as evidenced by experiments conducted on two open-source models.

Future work could investigate the effectiveness of alternative embedding and clustering methods in the proposed candidate selection phase. Additionally, since the demonstrations in our prompt were manually crafted and remained fixed across all inputs, future research could improve performance by exploring automatic example generation techniques or employing more advanced prompting strategies.

7 Limitations

This study is limited to experiments conducted with a single LLM (GPT-4o-mini) due to resource limitations. A broader evaluation involving multiple LLMs could offer deeper insights into the strengths and limitations of LLM-based post-processing methods for AAC.

References

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jee-weon Jung, Dong Zhang, HCH Yang, Shih-Lun Wu, David M Chan, Zhifeng Kong, D Ruifan, Z Yaqian, V Rafael, and Shinji Watanabe. 2024. Automatic audio captioning with encoder fusion, multi-layer aggregation, and large language model enriched summarization. Technical report, DCASE Challenge, Tech. Rep.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. 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 119–132.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Étienne Labbé, Thomas Pellegrini, and Julien Pinquier. 2024. CoNeTTE: An efficient audio captioning system leveraging multiple datasets with task embedding. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:3785–3794.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jizhong Liu, Gang Li, Junbo Zhang, Chenyu Liu, Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Yujun Wang, and Bin Wang. 2024. Leveraging CED encoder and large language models for automated audio captioning. Technical report, DCASE Challenge, Tech. Rep.

Chaitanya Prasad Narisetty, Tomoki Hayashi, Ryunosuke Ishizaki, Shinji Watanabe, and Kazuya Takeda. 2021. Leveraging state-of-the-art ASR techniques to audio captioning. In *Proc. Conf. Detection Classification Acoust. Scenes Events*, pages 160–164.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629, Miami, Florida, USA. Association for Computational Linguistics.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Mengyue Wu, Heinrich Dinkel, and Kai Yu. 2019. Audio caption: Listen and tell. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 830–834. IEEE.

Shih-Lun Wu, Xuankai Chang, Gordon Wichern, Jeeweon Jung, François Germain, Jonathan Le Roux, and Shinji Watanabe. 2024. Improving audio captioning models with fine-grained audio features, text embedding supervision, and llm mix-up augmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 316–320. IEEE.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kai Yu. 2022. The SJTU system for DCASE2022 challenge task 6: Audio captioning with audio-text retrieval pretraining. *Tech. Rep., DCASE2022 Challenge*.

Zhongjie Ye, Yuexian Zou, Fan Cui, and Yujun Wang. 2022. Automated audio captioning with multi-task learning. In *Proc. Conf. Detection Classification Acoust. Scenes Events*, pages 1–3.

Yiming Zhang, Hong Yu, Ruoyi Du, Zheng-Hua Tan, Wenwu Wang, Zhanyu Ma, and Yuan Dong. 2023. ACTUAL: Audio captioning with caption feature space regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zelin Zhou, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kenny Q. Zhu. 2022. Can audio captions be evaluated with image caption metrics? In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 981–985.

Memory-Efficient Training for Text-Dependent SV with Independent Pre-trained Models

Seyed Ali Farokh
Department of Computer Engineering
Amirkabir University of Technology
Tehran, Iran
alifarokh@aut.ac.ir

Hossein Zeinali
Department of Computer Engineering
Amirkabir University of Technology
Tehran, Iran
hzeinali@aut.ac.ir

Abstract

This paper presents our submission to the Iranian division of the Text-Dependent Speaker Verification Challenge (TdSV) 2024. Conventional TdSV approaches typically jointly model speaker and linguistic features, requiring unsegmented inputs during training and incurring high computational costs. Additionally, these methods often fine-tune large-scale pre-trained speaker embedding models on the target domain dataset, which may compromise the pre-trained models' original ability to capture speaker-specific characteristics. To overcome these limitations, we employ a TdSV system that utilizes two pre-trained models independently and demonstrate that, by leveraging pre-trained models with targeted domain adaptation, competitive results can be achieved while avoiding the substantial computational costs associated with joint fine-tuning on unsegmented inputs in conventional approaches. Our best system reached a MinDCF of 0.0358 on the evaluation subset and secured first place in the challenge.

Keywords: Text-dependent Speaker Verification, Speaker Verification, Memory-efficient Training, Pre-trained Models, Transfer Learning

1 Introduction

Speaker verification (SV) is the task of confirming an individual's identity based on their voice. It involves comparing one or more enrollment utterances with a test utterance and can be performed in either a text-independent (TiSV) or text-dependent (TdSV) setting. In TiSV, the phonetic content of the utterances is unrestricted, and only the speaker's identity is verified, whereas in TdSV, the system verifies both the speaker's identity and the specific phrase spoken. With the development of various neural network architectures (Xie et al., 2019; Desplanques et al., 2020; Zeinali et al., 2019b; Snyder et al., 2018), loss functions (Xiang et al., 2019; Zhang and Koishida, 2017; Wang et al., 2018; Deng et al., 2019), and pooling methods (Snyder et al., 2018; India et al., 2019; Zhu et al., 2018), TiSV has seen considerable improvement in recent years, whereas TdSV has remained relatively underexplored. TdSV systems can be either phrase-dependent (i.e., shared passphrases), where a fixed set of phrases is predefined by the system, or phrase-independent (i.e., user-defined passphrases), allowing users to customize their phrases (Hossein et al., 2024). With the growing demand for voice-based authentication, TdSV has gained increasing attention, as the phonetic content can be used as passphrases (Tu et al., 2022), adding an extra layer of security to voice-based access control systems.

This paper presents our system submitted to Task 1 of the Text-dependent Speaker Verification Challenge 2024[1] (Zeinali et al., 2025), which aimed to encourage participants to explore novel approaches for TdSV. The challenge was organized into two divisions: an international one, which included two subtasks focusing on shared and user-defined passphrases, and an Iranian division, which mirrored Task 1 of the worldwide challenge but specifically emphasized developing methods with limited GPU resources. In this challenge, model enrollment is done using three enrollment utterances, and each trial consists of a test utterance and a model identifier. Speaker verification trials fall into one of the following categories:

- **Target Correct (TC)**: The speaker matches the claimed model and utters the correct phrase.
- **Target Wrong (TW)**: The speaker matches the claimed model but utters an incorrect phrase.

[1]Challenge website: https://tdsvc.github.io

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 95–102
November 20-22, 2025 ©2025 Association for Computational Linguistics

- **Impostor Correct (IC)**: The speaker does not match the claimed model but utters the correct phrase.
- **Impostor Wrong (IW)**: The speaker does not match the claimed model and utters an incorrect phrase. This category was excluded from the current year's challenge, as it does not pose sufficient difficulty for contemporary models.

In the context of TdSV, proposed systems are required to integrate both speaker and phrase verification scores and accept only TC trials[2]. Task 1 is phrase-dependent, employing a fixed set of ten phrases (five in Persian and five in English) for enrollment and testing. Additionally, to enhance the complexity of the challenge, some test utterances in TW trials were sourced from free-text recordings.

The primary evaluation metric adopted by TdSV 2024 is the normalized minimum Detection Cost Function (MinDCF), as defined in NIST SRE 2008 as a weighted sum of miss and false error probabilities, with $P_{target} = 0.01$, $C_{FalseAlarm} = 1$, and $C_{Miss} = 10$. The Equal Error Rate (EER) will also be reported as a secondary performance measure.

Previous successful approaches to TdSV typically jointly model speaker characteristics and the linguistic content of utterances. For instance, Liu et al. (2021) proposed a phoneme-aware attentive pooling method that incorporates frame-level phoneme posteriors into attentive pooling, improving the model's ability to utilize phonetic information effectively. Also, some studies have employed supervised multi-task learning to jointly learn speaker and linguistic features for further improvement (Yang et al., 2020; Han et al., 2021).

However, joint speaker and phrase modeling has some drawbacks compared to independent modeling. First, model development becomes more complex than developing the system based on independent phrase and speaker embedding models. Additionally, since phrase modeling requires attending to an entire utterance, inputs cannot be chunked during training, requiring variable-length inputs to be zero-padded. This issue substantially increases GPU memory requirements, particularly for recent transformer-based models, due to their quadratic time and memory complexity (Vaswani et al., 2017).

[2]For Text-independent Speaker Verification (TiSV), the task definition differs: both TC and TW trials are accepted.

Furthermore, as demonstrated in this work, pre-trained speaker embedding models are highly effective at extracting speaker-related features while disregarding other information in input utterances. However, when subjected to multi-task fine-tuning, these models are prone to lose their initial ability to extract speaker-related features, allocating capacity to learning linguistic content instead. This shift reduces their effectiveness, especially when in-domain data for multi-task fine-tuning is limited.

Motivated by these challenges, we leverage the full capacity of pre-trained models and develop a TdSV system based on independent pre-trained models for phrase and speaker verification. For phrase verification, we fine-tune a pre-trained cross-lingual speech representation model for bilingual automatic speech recognition (ASR) in Persian and English, followed by a further fine-tuning stage for phrase classification. This classifier is used to reject incorrect phrases. Similarly, we develop several speaker embedding extractors based on pre-trained ResNets and Whisper (Radford et al., 2023) for our speaker verification system. After rejecting incorrect phrases using the phrase classifier, final verification scores are obtained by computing cosine similarity between test and enrollment embeddings.

Experimental results demonstrate that with well-designed fine-tuning stages, our TdSV system built on independently pre-trained models can achieve performance comparable to systems that jointly model speaker-related and linguistic information while using only a single Nvidia RTX 3090 GPU. This strategy substantially lowers GPU memory requirements and, consequently, reduces computational costs compared to the multi-GPU setups typically employed for training speaker recognition models (Zheng et al., 2023). Our best system secured first place in the Iranian division of the challenge and outperformed the third-place team in the international division (Zeinali et al., 2025).

The rest of the paper is organized as follows: Section 2 introduces the datasets used in this work. Sections 3 and 4 describe the architecture of our phrase and speaker verification systems, respectively. The experimental results and discussion are given in Section 5, and we conclude in Section 6.

2 Challenge Datasets

The DeepMine dataset (Zeinali et al., 2018, 2019a) is the primary source of the training and evaluation data for TdSV 2024. It was collected through crowd-sourcing, and while all participants were native Persian speakers, most contributed to the English portion of the dataset as well. The official TdSV 2024 data for Task 1 includes three subsets: training, development, and evaluation. The training subset consists of 183,431 utterances from 1,620 speakers. Among the utterances, 31,738 are free-text, while the rest were drawn from a fixed set of ten phrases comprising five Persian and five English phrases. The development and evaluation subsets are intended solely for system evaluation and contain 117,348 and 6,464,241 trials, respectively. During evaluation, model enrollment is conducted using three recordings of a specific phrase, and each trial includes a test utterance and a model identifier. The development set is provided to participants for evaluation and parameter tuning before submitting results to the official leaderboard. The evaluation subset is used for the official evaluation of the challenge. In addition to the DeepMine dataset, participants are also allowed to use the following datasets:

- **VoxCeleb 1&2** (Nagrani et al., 2017; Chung et al., 2018) are two large-scale datasets collected from YouTube videos, which contain over one million recordings from 7,205 celebrities. In this work, due to resource constraints, only VoxCeleb 1 was used, which includes over 100,000 utterances from 1,251 speakers.

- **LibriSpeech** (Panayotov et al., 2015) is a standard ASR corpus in US English that comprises approximately 1,000 hours of speech from 2,338 speakers. We only used the *train-clean-100* subset of this dataset to train our phrase verification system, which contains about 100 hours of speech.

- **Common Voice** (Ardila et al., 2020) is a multilingual speech dataset created from contributions of volunteers from worldwide. For this challenge, teams are restricted to using the Persian (Farsi) subset, which contains approximately 363 hours of validated speech from 4,148 speakers[3]. To prepare this subset for training our speaker verification systems, we excluded speakers with fewer than 30 recordings. From the remaining speakers with more than 650 recordings, we randomly selected 650 utterances per speaker, resulting in a final dataset with 125,017 utterances from 813 speakers.

The challenge rules prohibit the use of any other public or private data for training.

2.1 Data Augmentation

We did not use any augmentation methods in our phrase verification system. However, following the previous successful studies on speaker verification (Chen et al., 2022; Zheng et al., 2023), we adopted SoX-based speed perturbation by factors of 0.9 and 1.1 to triple the number of speakers during training, followed by an on-the-fly implementation of the following augmentations, each applied with a probability of 0.6: noise addition using the MUSAN dataset (Snyder et al., 2015), reverberation using RIRs dataset (Ko et al., 2017), and gain augmentation.

3 Phrase Verification System

Our proposed system for TdSV 2024 consists of two independent subsystems for phrase and speaker verification. The phrase verification system is a classifier that rejects TW trials, while the speaker verification system is responsible for producing similarity scores. Although this system design does not benefit from joint modeling of speaker and text, it greatly simplifies the system development process and allows for the use of various pre-trained models for each subsystem with minimal modifications.

The phrase classifier is an 11-class model trained with standard softmax. The first ten classes correspond to the set of phrases in the challenge, and the final class represents free text (or "none of the above"). This classifier is built on XLSR[4] (Conneau et al., 2021), a pre-trained cross-lingual speech representation model trained by solving a self-supervised contrastive task, proven to be effective in low-resource languages compared to traditional feature extraction methods. This model takes a raw waveform as input and produces a sequence of features.

Moreover, to improve the model's ability to extract linguistic features from Persian and English inputs, we initially fine-tuned the XLSR for bilingual speech recognition in Persian and English.

[3] Common Voice 18.0, released on 6/19/2024

[4] Facebook/wav2vec2-xls-r-300m

System	Full Training			Domain Adaptation		
	Epoch	BS	LR	Epoch	BS	LR
S2	-	-	-	15	32	3e-4
S3	100	64	1e-3	15	32	3e-4
S4	15	64	1e-3	7	28	5e-5
S5	15	64	1e-3	7	28	5e-5

Table 1: Hyper-parameters used in different submitted systems S2–S5 (BS = batch size, LR = learning rate).

During this phase, 30% of the training subset of Common Voice Farsi and LibriSpeech (*train-clean-100*) were used, and the model was trained using CTC loss (Graves et al., 2006) for 40 epochs, with an initial learning rate of 0.001 and an effective batch size of 32. In our experiments, this phase contributes to improving the performance of the phrase verification system.

Finally, to train the classifier, an attention-based pooling layer was added to the fine-tuned XLSR to compute fixed-dimensional utterance-level feature vectors from frame-level representations h_t $(t = 1, ..., T)$:

$$e_t = W_1 h_t + b_1, \tag{1}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{\tau}^{T} \exp(e_\tau)}, \tag{2}$$

$$\tilde{h} = \sum_{t}^{T} \alpha_t (W_2 h_t + b_2), \tag{3}$$

where, e_t and α_t are the attention score and weight, respectively. $\tilde{h}$ refers to the utterance-level feature vector, which is finally fed to a fully connected layer with ReLU activation, followed by a linear classifier. The network was trained using the Cross-Entropy loss function for one epoch on the entire training samples of the challenge dataset, with a learning rate of 0.0005 and an effective batch size of 64.

4 Speaker Verification System

To leverage the full power of pre-trained SV models and mitigate the computational cost of training randomly initialized models, we explored two directions for developing our SV system. In the first approach, we fine-tuned several pre-trained ResNet-based models, widely used as a standard architecture in speaker verification. In the second approach, we studied the performance of pre-trained ASR models adapted for SV, which have shown promising results in previous studies (Zhang et al., 2022; Cai et al., 2023; Liao et al., 2023). More specifically, we employed the Whisper-PMFA (Zhao et al., 2024) method, which involves fine-tuning a pre-trained Whisper model for speaker recognition.

4.1 Training Protocol

We trained our models in two stages:

- **Full training** (T_1): In this stage, models were trained on a combination of out-of-domain data (Common Voice Farsi and VoxCeleb 1) and in-domain (DeepMine) data, totaling 3,684 speakers, to learn robust and generalizable speaker embeddings across different domains. Pre-trained ResNets did not undergo this stage, as they are already capable of extracting rich speaker-specific features. During this phase, 300 consecutive frames of each input utterance were randomly selected in each epoch to prevent overfitting, reduce GPU memory usage, and accelerate training. Moreover, all augmentation methods explained in Section 2.1 were applied. We employed the widely used AAM-Softmax (Deng et al., 2019) loss with the subcenter method and the Inter-TopK penalty (Zhao et al., 2021) to train our models, with a constant margin and scale of 0.2 and 32, respectively.

- **Domain adaptation** (T_2): We fine-tuned our models using in-domain data after full training to bridge the domain gap and improve performance. During this stage, augmentation methods and the Inter-TopK penalty were removed to prevent domain mismatch. Additionally, the number of randomly selected frames was increased from 300 to 600 to enhance the models' generalization capability (Garcia-Romero et al., 2019, 2020). Fine-tuning was performed with smaller learning rates to preserve the models' generalization abilities.

All models were optimized using SGD with a momentum of 0.9 and a weight decay of 1e-4. We also utilized an exponential decay scheduler with a minimum learning rate of 5e-5 for T_1 and 1e-6 for T_2. Other training hyper-parameters are listed in Table 1. Note that gradient accumulation was used to achieve the target effective batch size when GPU memory was limited. The dimensionality of speaker embeddings was set to 256 across all models. All experiments were conducted on a sin-

System	Architecture	Training Stages	Development		Evaluation	
			$\text{MinDCF}_{0.01}$	EER(%)	$\text{MinDCF}_{0.01}$	EER(%)
S1	ResNet34		0.0614	1.3938	0.0784	1.7390
S2	ResNet293	T_2	0.0225	0.8733	**0.0376**	**1.1080**
S3	ResNet152	$T_1 + T_2$	0.0191	0.6757	0.0764	2.3444
S4	Whisper-PMFA	$T_1 + T_2$	0.0163	**0.6121**	0.0584	2.0410
S5	Whisper-PMFA	$T_1 + T_2$	**0.0161**	0.6126	0.0583	2.0445
Fusion (S1∼S5)			**0.0119**	**0.5605**	**0.0358**	1.2457

Table 2: Results of different submissions on the development and evaluation sets.

Subset	$\text{MinDCF}_{0.01}$	EER(%)
Development	0.0000	0.00
Evaluation	0.0003	0.01

Table 3: Phrase verification performance on TC-vs-TW trials.

gle Nvidia RTX 3090 GPU using the WeSpeaker toolkit (Wang et al., 2024).

4.2 ResNet

ResNet (Xie et al., 2019) is a widely used architecture for speaker recognition that has performed excellently in previous speaker verification challenges (Zheng et al., 2023). Consequently, many open-source implementations and pre-trained models have been publicly released based on this architecture. Trained on large-scale datasets like VoxCeleb 1&2, these pre-trained models can provide a robust starting point for training speaker recognition models on other datasets by improving their generalization and speeding up the convergence.

During the challenge period, we submitted three systems based on a bottleneck-block ResNet, all adopting temporal statistics pooling (Snyder et al., 2018) for aggregating variable-length sequence features into utterance-level embeddings. The first system (S1) was a pre-trained ResNet34 without domain adaptation, while the second one (S2) was a pre-trained ResNet293 that underwent domain adaptation. Finally, we applied both training stages to a randomly initialized ResNet152 to obtain our last ResNet-based system (S3).

4.3 Whisper-PMFA

Building on the successful use of pre-trained ASR models in speaker verification (Zhang et al., 2022; Cai et al., 2023), Zhao et al. (2024) recently proposed Whisper-PMFA (Partial Multi-Scale Feature Aggregation using Whisper) to leverage the capabilities of Whisper, a large-scale multilingual ASR model based on transformer architecture. Whisper-PMFA adapts Whisper for speaker verification by selectively concatenating frame-level outputs from specific transformer layers rather than aggregating features from all layers. This approach not only reduces computational overhead but also enhances performance by minimizing the integration of irrelevant information from lower-impact layers.

Inspired by this, we studied the performance of Whisper-PMFA in this challenge. Since Whisper was not trained for the speaker recognition task, we applied both training stages to Whisper-PMFA. Additionally, before the full training stage, we froze the Whisper parameters and fine-tuned the model for five epochs to prevent updating the pre-trained model in the wrong direction due to the random initialization of newly added components. We submitted two Whisper-PMFA-based systems (S4 and S5) to this challenge, differing only in the AAM-Softmax margin used during the domain adaptation phase: 0.35 for S4 and 0.2 for S5.

4.4 Feature Extraction

80-dimensional log Mel filter bank energies with a 25ms window and 10ms frame-shift were extracted for our ResNet-based models. Voice activity detection (VAD) was not applied, and all features were mean-normalized. Likewise, 80-dimensional log magnitude Mel spectrograms consistent with the pre-trained Whisper were utilized for training Whisper-PMFA.

4.5 Backend

Speaker embeddings were extracted from the final fully connected layer of the models, and cosine similarity was used to compute scores. Since model

Methods	Development	
	$\text{MinDCF}_{0.01}$	EER(%)
Whisper-PMFA (T_1)	0.0234	0.9253
+ Domain adaptation (T_2)	0.0177	0.6273
++ AS-Norm	0.0161	0.6126

Table 4: Ablation study on Whisper-PMFA.

enrollment is done using three utterances in this challenge, we used the average of embedding vectors of each model during scoring.

Afterward, AS-Norm (Wang et al., 2020) was used for score normalization, using 1,620 cohorts obtained from speaker-wise averaging of all embeddings in the training subset of the challenge dataset. The top 300 most similar scores were selected to compute the mean and standard deviation for normalization.

Finally, we adopted score fusion by averaging single-system scores to further improve performance.

5 Results

Table 2 shows the evaluation results of our single and fusion systems on the development and evaluation subsets of the challenge after applying AS-Norm and rejecting TW trials. The results indicate that the Whisper-PMFA method outperforms the widely used ResNet architecture with random initialization, conforming to the findings of previous studies on the effectiveness of adapting pre-trained ASR models for speaker verification. However, it can be observed from the results that the ResNets pre-trained on approximately twice the data (VoxCeleb 1&2) can considerably surpass Whisper-PMFA after a well-designed domain adaptation stage, which highlights the importance of large-scale pre-training in improving the generalization ability of speaker verification models.

In addition, Figure 1 presents the Detection Error Tradeoff (DET) curves of the best-performing system for different categories of evaluation data. The results indicate that the model generally performs better on Persian phrases, which is expected given that the DeepMine dataset was collected from native Persian speakers, many of whom are likely less fluent in English. Furthermore, the results show noticeably higher performance for male speakers compared to female speakers. This disparity is not solely due to the inherent challenges of verifying

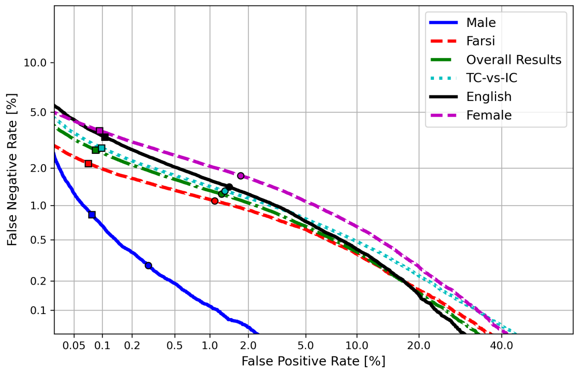


Figure 1: DET curves of our best-performing system.

female voices, but is also influenced by the specific characteristics of the DeepMine dataset, as discussed in its original description (Zeinali et al., 2018, 2019a) and in the official challenge results paper (Zeinali et al., 2025).

We also report the MinDCF and EER of the proposed phrase verification system on TC-vs-TW trials of the development and evaluation subsets (Table 3). According to the results, our phrase verification system demonstrates a near-optimal performance on this task.

5.1 Ablation Study

We conducted an ablation study on our Whisper-PMFA system (S5). The development set of the challenge dataset was used as our evaluation benchmark. We can observe from the results (Table 4) that the domain adaptation phase improved the MinDCF from 0.0234 to 0.0177. Also, a further improvement of MinDCF to 0.0161 was achieved after applying AS-Norm.

5.2 Comparison with Other Teams

To contextualize our performance, we report in Table 5 the evaluation results of our best system alongside the top-performing submissions in Task 1 of the international division of the TdSV Challenge. Team names and scores are taken directly from the official challenge results paper (Zeinali et al., 2025), which also provides brief descriptions and comparisons of the proposed architectures. As shown, our system achieves a lower MinDCF than the team ranked third in the international division.

6 Conclusion

In this paper, we present our system for Task 1 of the Iranian division of the Text-dependent Speaker Verification (TdSV) Challenge 2024, fo-

Team	MinDCF$_{0.01}$	EER(%)
Team 04 (Sreekanth, 2024)	**0.0297**	1.132
Team 08	0.0326	**1.013**
Our System	0.0358	1.246
Team 02	0.0379	1.164
Team 01	0.0504	2.245

Table 5: Evaluation results for our best system and the top-ranked teams in Task 1 of the international division of TdSV.

cusing on resource-constrained training for TdSV systems. Unlike previous methods that jointly model speaker-related and linguistic features, our approach leverages two independent pre-trained models for phrase and speaker verification. This design reduces the computational costs associated with joint modeling during training while fully utilizing the capabilities of pre-trained models to achieve competitive performance. Our best system achieved a MinDCF of 0.0358 on the evaluation subset, securing first place in the challenge.

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Danwei Cai, Weiqing Wang, Ming Li, Rui Xia, and Chuanzeng Huang. 2023. Pretraining Conformer with ASR for speaker verification. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zhengyang Chen, Bing Han, Xu Xiang, Houjun Huang, Bei Liu, and Yanmin Qian. 2022. SJTU-AISpeech system for VoxCeleb speaker recognition challenge 2022. *arXiv preprint arXiv:2209.09076*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep speaker recognition. In *Interspeech 2018*. ISCA.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pages 2426–2430.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, pages 3830–3834.

Daniel Garcia-Romero, Greg Sell, and Alan Mccree. 2020. MagNetO: X-vector magnitude estimation network plus offset for improved speaker recognition. In *The Speaker and Language Recognition Workshop (Odyssey 2020)*, pages 1–8.

Daniel Garcia-Romero, David Snyder, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur. 2019. X-Vector DNN refinement with full-length recordings for speaker recognition. In *Interspeech 2019*, pages 1493–1496.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Bing Han, Zhengyang Chen, Zhikai Zhou, and Yanmin Qian. 2021. The SJTU system for short-duration speaker verification challenge 2021. In *Interspeech 2021*, pages 2332–2336.

Zeinali Hossein, Lee Kong Aik, Alam Jahangir, and Burget Lukas. 2024. Text-dependent speaker verification (TdSV) challenge 2024: Challenge evaluation plan. *arXiv preprint arXiv:2404.13428*.

Miquel India, Pooyan Safari, and Javier Hernando. 2019. Self multi-head attention for speaker recognition. In *Interspeech 2019*, pages 4305–4309.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5220–5224. IEEE.

Dexin Liao, Tao Jiang, Feng Wang, Lin Li, and Qingyang Hong. 2023. Towards a unified Conformer structure: from ASR to ASV task. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yan Liu, Zheng Li, Lin Li, and Qingyang Hong. 2021. Phoneme-aware and channel-wise attentive learning for text dependent speaker verification. In *Interspeech 2021*, pages 101–105.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Sankala Sreekanth. 2024. Exploring self-supervised representations for text-dependent speaker verification. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1232–1239.

Youzhi Tu, Weiwei Lin, and Man-Wai Mak. 2022. A survey on text-dependent and text-independent speaker verification. *IEEE Access*, 10:99038–99049.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Shuai Wang, Zhengyang Chen, Bing Han, Hongji Wang, Chengdong Liang, Binbin Zhang, Xu Xiang, Wen Ding, Johan Rohdin, Anna Silnova, et al. 2024. Advancing speaker embedding learning: WeSpeaker toolkit for research and production. *Speech Communication*, 162:103104.

Weiqing Wang, Danwei Cai, Xiaoyi Qin, and Ming Li. 2020. The DKU-DukeECE systems for voxceleb speaker recognition challenge 2020. *arXiv preprint arXiv:2010.12731*.

Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu. 2019. Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1652–1656.

Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2019. Utterance-level aggregation for speaker recognition in the wild. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5791–5795. IEEE.

Yexin Yang, Shuai Wang, Xun Gong, Yanmin Qian, and Kai Yu. 2020. Text adaptation for speaker verification with speaker-text factorized embeddings. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6454–6458.

Hossein Zeinali, Lukáš Burget, and Jan Honza Černockỳ. 2019a. A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 397–402. IEEE.

Hossein Zeinali, Kong Aik Lee, Jahangir Alam, and Lukáš Burget. 2025. Text-dependent speaker verification challenge 2024: Exploring shared and user-defined passphrases. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Hossein Zeinali, Hossein Sameti, and Themos Stafylakis. 2018. DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in Persian and English. In *Odyssey*, pages 386–392.

Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot. 2019b. BUT system description to voxceleb speaker recognition challenge 2019. *arXiv preprint arXiv:1910.12592*.

Chunlei Zhang and Kazuhito Koishida. 2017. End-to-end text-independent speaker verification with triplet loss on short utterances. In *Interspeech 2017*, pages 1487–1491.

Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung yi Lee, and Helen Meng. 2022. MFA-Conformer: Multi-scale feature aggregation conformer for automatic speaker verification. In *Interspeech 2022*, pages 306–310.

Miao Zhao, Yufeng Ma, Min Liu, and Minqiang Xu. 2021. The SpeakIn system for voxceleb speaker recognition challange 2021. *arXiv preprint arXiv:2109.01989*.

Yiyang Zhao, Shuai Wang, Guangzhi Sun, Zehua Chen, Chao Zhang, Mingxing Xu, and Thomas Fang Zheng. 2024. Whisper-PMFA: Partial multi-scale feature aggregation for speaker verification using Whisper models. In *Interspeech 2024*, pages 2680–2684.

Yu Zheng, Yajun Zhang, Chuanying Niu, Yibin Zhan, Yanhua Long, and Dongxing Xu. 2023. Unisound system for voxceleb speaker recognition challenge 2023. *arXiv preprint arXiv:2308.12526*.

Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey. 2018. Self-attentive speaker embeddings for text-independent speaker verification. In *Interspeech 2018*, pages 3573–3577.

Information-theoretic conditioning in terminological alternations in specialized domains: The cases of Taiwan Mandarin legal language and English biomedical language

Po-Hsuan Huang
Department of Linguistics
University of Southern California
3601 Watt Way, Los Angeles, CA 90089
pohsuan@usc.edu

Hsuan-Lei Shao
Institute of Medical and
Biotechnology Law
Taipei Medical University
301 Yuantong Rd. Zhonghe Dist.
New Taipei City, Taiwan, 235603
hlshao@tmu.edu.tw

Abstract

This study examines how information-theoretic correlates, specifically contextual surprisal, condition terminological alternations in specialized domains, where both domain-specific and general terms express similar concepts. Specifically, two competing theories exist. The Uniform Information Density (UID) theory proposes that the speaker would avoid abrupt information rate changes. This predicts the use of more specific variants when the surprisals are higher. Conversely, availability-based production suggests the use of more readily-accessible items with higher surprisals. This study examines the dynamics between these two potential mechanisms in the terminological use in specialized domains. Specifically, we argue that, in specialized language, due to the higher frequency of domain-specific terms, both accounts predict the use of specific items in higher-surprisal contexts. The cases of Taiwan Mandarin legal language and English biomedical language were, therefore, examined. Crucially, a current popular method for probability estimation is through large language models (LLMs). The linguistic distribution in specialized domains, however, may deviate from the general linguistic distribution on which the LLMs are trained. Thus, we propose a novel semantics-based method of estimating the token probability distribution in a given corpus that avoids the potentially different linguistic distribution and the issue of word segmentation. As expected, results indicated a positive correlation between a variable's surprisal and the use of domain-specific variants in both cases. This supports UID-based production, and arguably also availability-based production, since more specific and frequent variants are preferred in high-surprisal contexts. Specifically, our semantics-based probability estimation outperformed LLM-based estimation and the baseline in both cases. This suggests the feasibility of semantics-based probability estimation in specialized domains.[1]

Keywords: domain-specific variation, information theory, surprisal calculation, semantics

1 Introduction

A growing number of studies have come to emphasize the role of information-theoretic (Shannon, 1948) constraints in communication and the conditioning of these constraints on linguistic distributions. Especially, lexical and syntactic production and processing are attested to be conditioned by information-theoretic correlates, including word frequency and contextual surprisal. Zhan and Levy (2018), for example, investigated the choice of classifiers in Mandarin and found that while frequency did not have an effect, the surprisal of the following noun could predict the language user's choice of classifiers. When the following noun had a higher contextual surprisal, the language user was more likely to opt for the general classifier *ge*, as opposed to the other specific classifiers. Likewise, in Wilcox et al.'s (2023) reading time study across 11 languages, it was found that both contextual surprisal and contextual entropy were positively correlated with the subjects' reading time.

[1] The code implementation of this study is available at: https://github.com/Peh-Suan/information_theoretic_conditioning_domain_specific.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 103–107
November 20-22, 2025 ©2025 Association for Computational Linguistics

1.1 Speaker-centric vs. listener-centric production

Importantly, two competing mechanisms have been put forth. The Uniform Information Density (UID) theory (Levy and Jaeger, 2007; Jaeger, 2010) proposes that during communication, the speaker would prevent abrupt information rate changes to facilitate better speech comprehension. Conversely, a more speaker-centric account, availability-based production (Bock, 1987; Ferreira and Dell, 2000), predicts that the speaker would prefer more readily accessible items. These two mechanisms, therefore, make opposite predictions: While UID would predict the use of more specific variants when the variable is contextually surprising, availability-based production would expect more general items to be used, as they are more accessible than specific ones.

In this study, we examine these two potential mechanisms in the terminological alternations in specialized domains. Specifically, several studies have suggested availability-based production in lexical-syntactic alternations. For example, Zhan and Levy (2018) examined how the contextual surprisal of a noun might influence the use of the general classifier *ge* vs. specific classifiers in Mandarin. It was found that when the noun had a higher surprisal, there was a higher tendency for the speaker to use the general classifier. Likewise, such availability-based production was also attested in Russian comparative constructions (Clark et al., 2022). In Russian, there are two options for comparative construction. The first one is the explicit option, where "than" is used. The other is the genitive option, where the target noun phrase being compared is marked with the genitive case, and "than" is omitted. In the first construction, there is an additional morpheme before going into the target noun phrase, while in the second construction, there is no such buffer. The first construction thus provides a higher availability for the speaker's speech planning. Indeed, it was also found that when the target noun phrase was more complex, the explicit option was preferred.

1.2 Terminological alternations in specialized domains

All the previous studies, however, focused on general language use. It therefore remains unknown whether style differences exist between general and domain-specific language.

Crucially, it is likely that both accounts may favor the domain-specific terms in high-surprisal contexts in domain-specific language. In specialized domains, the same concepts may be expressed through different terms. In English biomedical language, *dermis* or *epidermis* can be used instead of *skin*. Similarly, in Taiwan Mandarin legal language, *zhi.yan.zhi* "in sum" can be used instead of the more colloquial *jian.yan.zhi*.

In the general context, the general terms are without doubt more frequently used. In the specialized domains, however, the respective domain-specific terms may actually be more frequent than the general counterparts. Indeed, in the corpora in this study, the domain-specific terms are 2.18 and 3.27 times more frequent than the general terms in Taiwan Mandarin legal language and English biomedical language, respectively. This, therefore, suggests that both the availability-based production and UID may support the use of domain-specific items when the surprisals are higher.

Therefore, in this study, we examine the information-theoretic conditioning, specifically the effects of surprisal, on terminological alternations in Taiwan Mandarin legal language and English biomedical language.

2 Methods

To answer how contextual surprisal interacts with terminological alternations, two corpora were examined. The contextual surprisals of the terminological variables were calculated based on the popular LLM-based probability estimation and our proposed semantics-based estimation. Linear-mixed effects models were used for statistical analysis.

2.1 Corpora

2.1.1 Taiwan Mandarin legal corpus

The Taiwan Mandarin legal corpus was built from 383,733 legal judgments made in 2024 obtained from the Governmet OpenData platform (`http://data.gov.tw`). Sentence segmentation was performed based on punctuation. 580,593 sentences were collected. 100,000 sentences were then randomly selected as the final corpus.

2.1.2 English biomedical corpus

A subset of the PMC corpus (National Library of Medicine, 2024) was used to build the English biomedical corpus. 1,029,191 sentences were collected. 100,000 sentences were then randomly selected as the final corpus.

2.1.3 Terminological variable selection

The terminological variables were manually inspected and selected by the authors. Only variables with higher frequencies were included. 15 general-vs.-legal and 25 general-vs.-biomedical terminological variables were chosen. An example of such variables is the *skin* vs. *dermis/epidermis* alternation mentioned previously. In this example, *skin*, *dermis*, and *epidermis* are all variants of this variable.

2.2 Surprisal estimation

The contextual surprisal of a token w given the context c is $-logP(w|c)$. To calculate a token's contextual surprisal, therefore, its probability in the corpus has to be estimated.

A conventional method of calculating probability is to calculate the raw frequency of the token. This is, however, not ideal for contextual surprisal estimation, since the likelihood of the exact context sentence happening more than once is low.

A more popular alternative is to directly estimate $P(w|c)$ through trained large language models (LLMs). This, however, may also not be ideal since the style differences between general and specialized language may lead to different linguistic distributions.

Therefore, in this study, we propose a novel semantics-based probability estimation based on the "semantic bit count" instead of the raw frequency of the tokens. We propose that, since information-theoretic correlates are essentially based on the amount of information, the semantics of the word token could be more revealing than pure token counts.

2.2.1 Semantics-based probability estimation

In this study, we propose counting a token's semantic bit occurrences in the corpus to estimate the probability of the token. Given the word embedding of a token w, and the embedding of a context sentence c (calculated as the mean of all the token embeddings in the sentence), the number of semantic bits of w in c can be approximated as the cosine similarity between the two vectors.

To convert this cosine similarity to a semantic bit count (sb), it is then rescaled from -1 to 1 to 0 to 1. This semantic bit count is then used instead of the raw frequency. The final semantics-based probability estimation of a token w in the context c is shown in Eq. 1, where C is all the context sentences in the corpus and C_w is all the context sentences where w occurs.

$$\hat{P}_{semantics}(w|c) = \frac{\sum_{c_j \in C_w} sb(w, c_j)}{\sum_{c_i \in C} sb(c, c_i)} \quad (1)$$

To compare the performance of semantics-based surprisal ($I_{semantics}$), LLM-based[2] surprisal (I_{LLM}) and baseline surprisal ($I_{baseline}$), which were calculated with direct 5-gram context counts, were also calculated.

2.3 Statistical analysis

Logistic-mixed effects models (LMMs) were used to test statistical significance through Satterthwaite's method. A model was fitted for each of the three kinds of surprisals for each of the two corpora.

The use of general vs. domain-specific was contrast coded as -0.5 (general) and 0.5 (domain-specific). Surprisal was standardized and taken as the predictor. Standardized frequency was included as a control variable. Random intercepts were grouped by terminological variable.

To compare the performance of the three types of surprisals, Akaike Information Criterion (AIC) was also used to test the relative quality of the fitted models.

3 Results

3.1 Taiwan Mandarin legal language

For both $I_{semantics}$ and I_{LLM}, positive correlations between the use of the domain-specific variants and the variable's contextual surprisal were found ($I_{semantics}$: $\hat{\beta} = 4.37$; $p < 0.001$;

[2]LLAMA-2-7B were used.

I_{LLM}: $\hat{\beta} = 0.11$; $p = 0.03$). On the flip side, $I_{baseline}$ was found to have insignificant effects ($\hat{\beta} = -0.06$; $p = 0.10$).

Crucially, the $I_{semantics}$ model had the lowest AIC ($I_{semantics}$: 6164.80; I_{LLM}: 6185.45; $I_{baseline}$: 6623.24), suggesting it is the most ideal model among the three.

3.2 English biomedical language

Similar positive effects were found between $I_{semantics}$ and domain-specific vs. general term use ($\hat{\beta} = 0.27$; $p < 0.001$). However, a negative correlation between I_{LLM} and domain-specific vs. general term use was found ($\hat{\beta} = -1.51$; $p < 0.001$). On the other hand, $I_{baseline}$ was once again found to have insignificant effects ($\hat{\beta} = 0.02$; $p = 0.48$).

In terms of the model quality based on AIC, the $I_{semantics}$ model was once again suggested to be the most ideal ($I_{semantics}$: 654.47; I_{LLM}: 2100.33; $I_{baseline}$: 5789.62).

4 Discussion

4.1 Information-theoretic conditioning in specialized language and speaker vs. listener-centric production

The main focus of this study is to examine how the style differences between general language and specialized language may interplay with speaker vs. listener-centric production. As discussed in Section 1.1, the UID theory and availability-based production are put forth as two competing mechanisms in previous studies (Zhan and Levy, 2018; Clark et al., 2022). It is suggested that, from a listener-centric perspective, the UID theory would predict more specific language use when the variable is more unpredictable/informative, in order to reduce abrupt information rate changes. From a speaker-based angle, on the other hand, the speaker would prefer more readily accessible variants. This would thus predict the use of more general items, which are presumably more accessible, when the unpredictability is higher.

These studies, however, focus on general language use. We argue that while such competition may hold in general/colloquial language, the two mechanisms may be compatible in specialized language. This is because the domain-specific terms may in fact be more frequent, and thus more accessible, than the general terms in specialized domains. Thus, both accounts would predict the use of domain-specific terms in higher-surprisal contexts, since they are at the same time more informative and readily accessible.

Indeed, the results in this study support our hypothesis. Positive correlations were attested between the semantics-based surprisal and the use of domain-specific terms in both cases. Indeed, opposite effects were found for the LLM-based surprisal, and no effects were found for the baseline surprisal. We argue, however, as will be discussed in the next section, that the semantics-based surprisal is the more appropriate estimation.

4.2 Semantics-based probability estimation for specialized language

The other contribution of this study is the proposal of a novel semantics-based probability estimation for specialized language. As argued in Section 2.2, contextual surprisal cannot be ideally calculated through raw token frequency, nor is it appropriate to use pretrained LLMs, as the linguistic distributions of specialized language may differ from that of general language.

In this study, we propose that the semantics of a token may be more information-theoretically relevant than pure occurrence frequency. A probability estimation based on the semantic bit count of the token was proposed. It was found that in both test cases, our method outperformed the LLM-based method and the baseline. Our results, therefore, suggest the feasibility of semantics-based probability estimation for specialized language in future studies.

Limitations

While the examination of competing surprisal candidates, i.e., the LLM-based surprisal and the baseline, allowed for a general investigation of the performance of our semantics-based method, it remains possible that different LLMs may lead to better or worse performances. In this study, we only examined one LLM (LLAMA-2-7B). To make the findings more grounded, a comparison between our method with a wider array of models may be ideal.

References

Kathryn Bock. 1987. An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language*, 26(2):119–137.

Thomas Hikaru Clark, Ethan Gotlieb Wilcox, Edward Gibson, and Roger P. Levy. 2022. Evidence for availability effects on speaker choice in the Russian comparative alternation. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*.

Victor S. Ferreira and Gary S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40(4):296–340.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 849–856.

National Library of Medicine. 2024. Pubmed central open access subset. `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`. Accessed: 2024-11-01.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Meilin Zhan and Roger Levy. 2018. Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1997–2005, New Orleans, Louisiana. Association for Computational Linguistics.

Voice Spoofing Detection via Speech Rule Generation Using wav2vec 2.0-Based Attention

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

Yu-Chen Gao
Department of Electrical Engineering,
Southern Taiwan University of Science and Technology
4B12C100@office.stust.edu.tw

Yu-Ying Xiao
Department of Electrical Engineering,
Southern Taiwan University of Science and Technology
4B127052@office.stust.edu.tw

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

Kun-Yi Huang
Department of Computer Science and Information Engineering, Southern Taiwan University of Science and Technology
iamkyh77@stust.edu.tw

Abstract

Recent advancements in AI-based voice cloning have led to increasingly convincing synthetic speech, posing significant threats to speaker verification systems. In this paper, we propose a novel voice spoofing detection method that integrates acoustic feature variations with attention mechanisms derived from wav2vec 2.0 representations. Unlike prior approaches that directly utilize wav2vec 2.0 features as model inputs, the proposed method leverages wav2vec 2.0 features to construct speech rules characteristic of bona-fide speech. Experimental results indicate that the proposed RULE-AASIST-L system significantly outperforms the baseline systems on the ASVspoof 2019 LA evaluation set, achieving a 24.6% relative reduction in equal error rate (EER) and an 10.8% reduction in minimum tandem detection cost function (min t-DCF). Ablation studies further confirm the importance of incorporating speech rules and selecting appropriate hidden layer representations. These findings highlight the potential of using self-supervised representations to guide rule-based modeling for robust spoofing detection.

Keywords: Voice spoofing detection, Speech rule generation, wav2vec 2.0

1 Introduction

Telecom fraud has become a critically important issue today, particularly the method of using AI to synthesize the voices of victims' family members to impersonate them and commit financial fraud. This has emerged as a new tactic employed by scam groups. Voice spoofing can be primarily divided into two categories: Physical Access (PA) attacks and Logical Access (LA) attacks. According to past research in relevant literature, the difficulty of signal detection in LA attacks is typically greater than that in PA attacks. This is primarily because voice conversion and text-to-speech technologies can more accurately mimic the target speaker's voice, rather than merely reproducing recorded playback quality.

To address the growing threat of voice spoofing attacks, many studies have adopted deep neural network (DNN)-based models to classify speech as either genuine or spoofed (Y. Zhang et al., 2021; J. Zhou et al., 2022; A. Gomez-Alanis et al., 2019). However, these approaches typically treat spoofing detection as a binary classification problem that focuses solely on surface-level acoustic differences,

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 108–115
November 20-22, 2025 ©2025 Association for Computational Linguistics

without accounting for the complexity and diversity of feature variations introduced by different spoofing methods. To address this limitation, (J. Boyd et al., 2023) proposed using a multi-class classification framework that distinguishes between genuine, voice conversion, speech synthesis, and replay categories. This enables the model to learn more discriminative features for identifying various types of spoofing attacks targeting genuine speech. However, most existing research on voice spoofing detection focuses on feature analysis from a single audio perspective.

Self-supervised learning (SSL) has emerged as a powerful alternative for extracting high-dimensional representations of speech signals (A. Baevski et al, 2020; W.-N. Hsu et al., 2021; S. Chen et al., 2022). These models typically rely on convolutional neural network (CNN)-based feature encoders, where CNN kernels perform nonlinear transformations on short segments of audio. A key advantage of self-supervised learning lies in its ability to learn from large-scale unlabeled data, enabling pre-trained models to capture a wide range of speech variability. Compared to conventional frequency-domain methods, these learned representations often yield more robust and informative features. Recently, SSL models such as wav2vec 2.0 (A. Baevski et al., 2020) have gained significant attention in various speech-related tasks. Originally developed for automatic speech recognition (ASR) (A. Bawitlung et al., 2025), these models have also demonstrated strong performance in speaker verification (Z. Fan et al., 2021) and speech emotion recognition (B. Nasersharif and M. Namvarpour, 2024). Recently, several studies have investigated the application of wav2vec 2.0 for spoofing detection tasks (H. Tak et al., 2022), taking advantage of its rich contextualized speech representations to improve feature modeling and detection accuracy.

This work proposes a novel framework that integrates conventional acoustic feature analysis with the sequential representation patterns derived from wav2vec 2.0. By exploring the interactions between acoustic features and the sequential rule of wav2vec 2.0 representations, the proposed approach enables voice spoofing detection not only from the inherent characteristics of speech but also through identifying inconsistencies in the sequence patterns of wav2vec 2.0 representations correlated with spoofed audio. This joint analysis enhances detection performance by uncovering unnatural patterns indicative of spoofed speech.

This paper addresses the fraudulent methods arising from current AI voice cloning technologies by proposing a detection method that combines the correlation between acoustic features and wav2vec 2.0-based attention mechanisms. This approach simultaneously considers the interaction between variations in acoustic features, and the rules of speech representations, aiming to enhance the accuracy of distinguishing between synthetic and genuine voices.

2 Related Work

2.1 AASIST

The AASIST network is composed of four main components: an encoder module, graph modules, a max graph operation (MGO) module, and an output module, as shown in the upper part of Figure 1. The encoder, based on RawGAT-ST (H. Tak et al., 2021), extracts high-level feature representations F directly from the raw audio waveform. Two parallel graph modules are employed to model the spectral and temporal characteristics of F, respectively, producing graph-structured features in both domains. These outputs are then fused to construct a heterogeneous graph, which is further processed by the MGO module.

The MGO module consists of two parallel upper and lower branches, each comprising two heterogeneous attention mechanisms and two stacked nodes that store time-frequency heterogeneous information. The final representation is obtained by applying an element-wise max operation to the outputs of the two branches. This representation is used to discriminate between bona-fide and spoofed speech.

2.2 wav2vec 2.0 Representations

wav2vec 2.0 leverages self-supervised learning to derive informative and high-level speech representations directly from raw audio input. Its architecture consists of two primary components: a convolutional feature extractor and a Transformer-based contextual module. The convolutional encoder transforms the input waveform into a sequence of latent vectors that capture fine-grained acoustic details. These latent features are subsequently processed by the contextual module, which employs self-attention mechanisms to

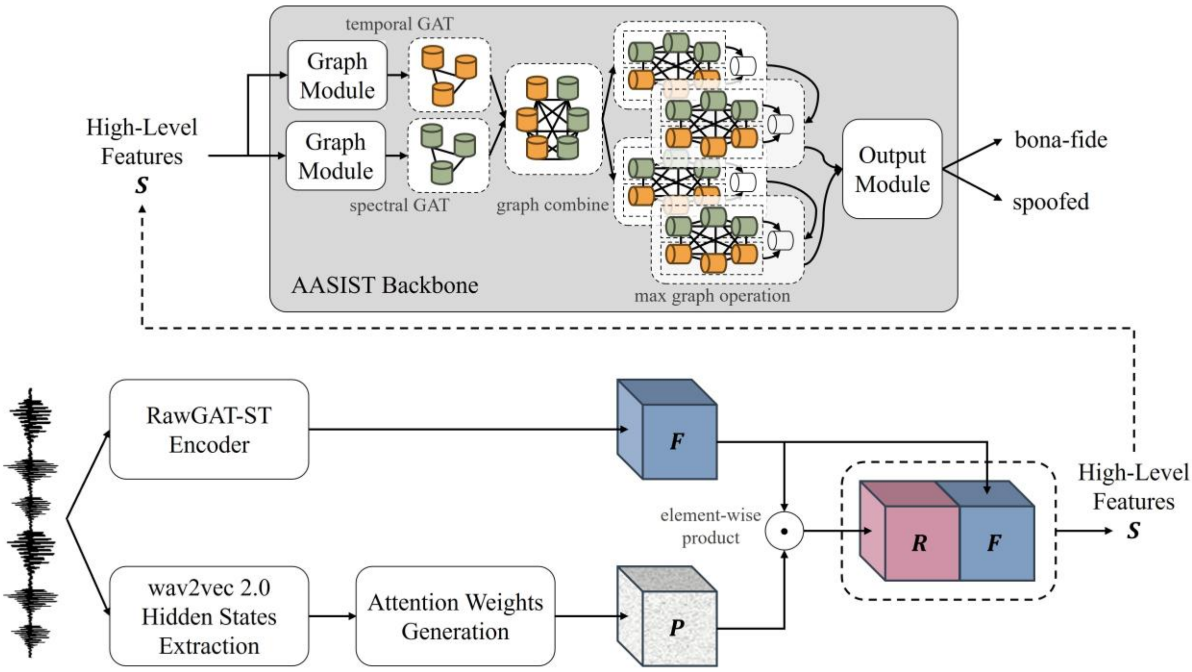


Figure 1: The proposed wav2vec 2.0-based attention network for high-level feature extraction in voice spoofing detection.

model temporal dependencies across the sequence, resulting in contextualized embeddings that reflect both short- and long-range speech characteristics. The model is pretrained using a contrastive objective, where segments of the latent sequence are masked and the network learns to distinguish the true representation from a set of distractors based on surrounding context. This training strategy enables wav2vec 2.0 to acquire phonetic and semantic knowledge from unlabeled speech data, making the learned representations broadly applicable to downstream tasks such as automatic speech recognition, speaker verification, and spoofing detection.

3 Speech Rule Generation via wav2vec 2.0-Based Attention

Unlike previous studies that directly utilize wav2vec 2.0 features as input to classification models, this work explores the use of wav2vec 2.0 representations to learn the underlying speech rules present in bona-fide speech. We hypothesize that spoofed speech introduces inconsistencies or deviations from these learned regularities. By identifying such rule violations, the proposed approach aims to enhance the accuracy of voice spoofing detection. The proposed wav2vec 2.0-based attention network for high-level feature extraction as depicted in Figure 1.

3.1 wav2vec 2.0-Based Attention

Initially, wav2vec 2.0 is used to extract hidden states s from the raw training audio, where $s \in \mathbb{R}^{T \times L}$ denotes a sequence of T time steps, each represented by an L-dimensional feature vector. These representations are then passed through N Residual Blocks for feature transformation. Assuming the $p_0 = s$, the standard transformation performed by each Residual Block is defined as follows:

$$p_{\mathrm{i}+1} = \mathcal{F}(\mathcal{F}(p_{\mathrm{i}}; \mathcal{K}_{i1}); \mathcal{K}_{i2}) + p_{\mathrm{i}} \quad (1)$$

where $\mathcal{F}(\cdot)$ denotes a 2D convolutional layer parameterized by kernel $\mathcal{K}_{(\cdot)}$, and each input of $\mathcal{F}(\cdot)$ is nonlinearly transformed by a composite function consisting of batch normalization followed by the scaled exponential linear unit (SELU) activation.

After that, the feature size of p_{N} and the raw audio after encoder processing are different, we

Layer	Input shape	Output shape
Raw audio	-	(64600)
Wa2vec 2.0 hidden states	(64600)	(199, 768)
Expand dim	(199, 768)	(1, 199, 768)
ResBlock_A × 2	(1, 199, 768)	(C_1, 199, 768)
ResBlock_B × 4	(C_1, 199, 768)	(C_2, 199, 768)
Conv2D	(C_2, 199, 768) kernel: (7, 11) stride: (7, 11)	(C_2, 29, 69)
BN	-	-
MaxPool	(C_2, 29, 69) kernel: (1, 3)	(C_2, 29, 23)
Softmax	dim=1	(C_2, 29, 23)
Hybrid High-Level Features	(C_2, 29, 23) combine R, F	(C_2, 58, 23)

Table 1: The speech rule generation architecture for voice spoofing detection.

apply local convolution and max pooling to compress the size of p_{N} to match the encoder output size.

$$\mathcal{T} = \mathrm{MaxPool}(\mathrm{BN}(\mathcal{F}(p_{\mathrm{N}}))) \tag{2}$$

$$P_{c,t,f} = \frac{e^{\mathcal{T}_{c,t,f}}}{\sum_{\tau=1}^{T} e^{\mathcal{T}_{c,\tau,f}}} \tag{3}$$

where $\mathrm{MaxPool}(\cdot)$ is max pooling, $\mathrm{BN}(\cdot)$ is batch normalization, and $P \in \mathbb{R}^{C_2 \times T \times F}$ can be defined as the attention weights employed to regulate speech rules.

3.2 Hybrid High-Level Features

Since P is the attention weights derived from the hidden states of wav2vec 2.0, we further apply an element-wise product between P and the encoder output F to generate the corresponding speech rules.

$$R = P \odot F \tag{4}$$

Finally, the speech rule R is used as auxiliary features and concatenated with F to obtain the hybrid high-level features. The size for each layer is illustrated in Table 1.

4 Experimental Results

4.1 Data Preparation

In alignment with the data preparation methodology outlined in (J.-w. Jung et al., 2022), all experiments in this study are conducted using the LA partition of the ASVspoof 2019 dataset (M. Todisco et al., 2019). The dataset is divided into three distinct subsets: training, development, and evaluation. The training and development subsets include spoofed speech generated using six known attack algorithms (A01–A06), while the evaluation subset extends this with an additional set of seven attack methods (A07–A19). Furthermore, the ASVspoof 2021 (J. Yamagishi et al., 2021) evaluation set is used to evaluate the cross-corpus performance of the proposed voice spoofing detection method.

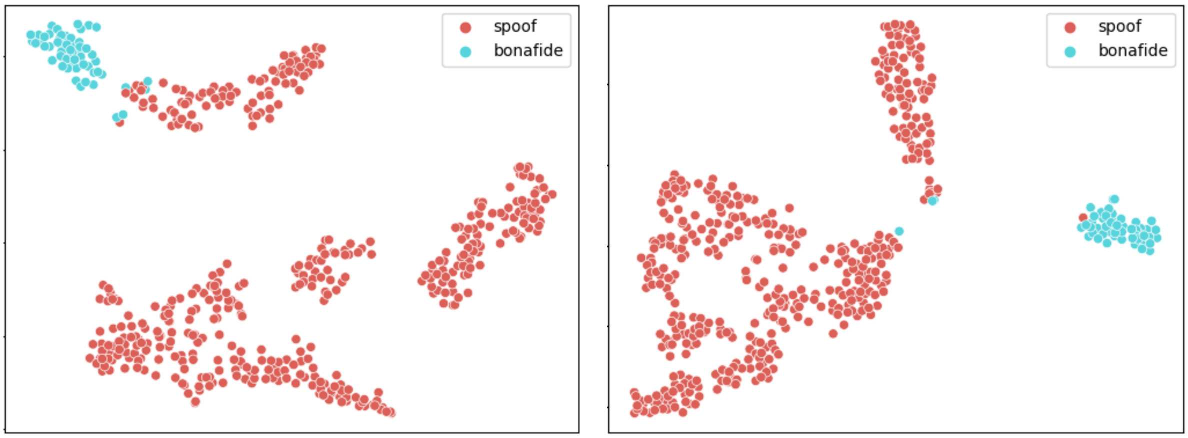
In this paper, we employ the “facebook/wav2vec2-base-960h” model, a Transformer-based architecture designed for speech representation learning. The model is pretrained in a self-supervised manner on 960 hours of unlabelled audio from the LibriSpeech corpus and later fine-tuned for automatic speech recognition tasks. Its structure consists of a convolutional feature extractor followed by twelve Transformer encoder layers, enabling the model to capture hierarchical representations of speech signals. We extract hidden states from both intermediate layers and the final layer. The intermediate layers are known to preserve more acoustic-level and phonetic information, making them well-suited for tasks that require detailed speech characteristics such as prosody, speaker traits, or subtle temporal variations. In contrast, the final layer tends to encode high-level semantic features aligned with the ASR objective, capturing more abstract linguistic content but potentially discarding lower-level acoustic cues.

4.2 Experimental Setup

In our experiments, we adopt lightweight variant AASIST-L as the backbone architecture, following the experimental setup outlined in (J.-w. Jung et al., 2022). The input to the model consists of raw audio waveforms with a fixed length of 64,600 samples, corresponding to approximately four seconds of speech. No data augmentation techniques are applied during training, ensuring that all models are trained on the original waveform data without synthetic variation. Model training is conducted

System	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	EER	min t-DCF
RawNet2 [13]	9.8	17.9	7.3	8.9	4.2	8.8	2.0	1.3	7.3	4.6	2.4	62.9	5.8	5.54	0.1547
RawGAT-ST [14]	1.19	0.33	0.03	1.54	0.41	1.54	0.14	0.14	1.03	0.67	1.44	3.22	**0.62**	1.19	0.0333
AASIST-L (reproduced)	**0.45**	0.34	0.02	**0.63**	0.34	**0.69**	0.19	0.23	0.53	**0.42**	1.96	2.97	0.88	1.14	0.0316
RULE-AASIST-L (proposed)	0.77	**0.16**	**0.02**	0.90	**0.16**	0.79	**0.12**	**0.10**	**0.42**	0.57	**1.18**	**2.34**	0.87	**0.86**	**0.0282**

Table 2: EER (%) and minimum t-DCF results for baseline and proposed model on the ASVspoof 2019 LA evaluation set.



(A) AASIST-L (B) RULE-AASIST-L

Figure 2: The distribution of output features from the last hidden layer of different models visualized using t-SNE, based on 240 randomly selected samples.

using the Adam optimizer with a batch size of 24 and a total of 100 training epochs. The objective function used is categorical cross-entropy loss.

As demonstrated in the study by (X. Wang and J. Yamagishi, 2021), the performance of spoofing detection systems can vary considerably depending on the choice of random seed due to the inherent stochasticity of the training process. To ensure a fair and robust evaluation, all experiments in this work are conducted using three different random seeds. In the experimental analysis, this paper reports the best result obtained from model training conducted with three different random seeds.

To evaluate system performance, we adopt two widely used metrics: the minimum tandem detection cost function (min t-DCF) and the equal error rate (EER).

4.3 Voice Spoofing Detection Results

The results are summarized in Table 2. Compared to the baseline systems, the proposed RULE-AASIST-L demonstrates significantly improved performance. Under the same backbone architecture and experimental setup, RULE-AASIST-L achieves a relative improvement of 24.6% in EER (i.e., 0.86% vs. 1.14%) and an 10.8% reduction in min t-DCF (i.e., 0.0282 vs. 0.0316), highlighting the effectiveness of the proposed method. On the other hand, Figure 2 illustrates the distribution of the output features from the last hidden layer of different models. It is evident that the proposed RULE-AASIST-L model yields more compact distributions for both spoof and bonafide classes compared to the AASIST-L baseline. This indicates that the RULE-AASIST-L model can more effectively distinguish between genuine and spoofed speech.

The results indicate that RULE-AASIST-L successfully leverages the attention generated during model training to define bona-fide speech rules. These learned rules help identify inconsistencies in spoofed speech, thereby

System	EER	min t-DCF
RULE-AASIST-L	**0.86**	**0.0282**
w/o ***F*** in the high-level features ***S***	1.54	0.0468
Use only ***P*** as the high-level features ***S***	2.80	0.0830
Replace wav2vec 2.0 hidden state extraction from layer 6 with layer 12	1.29	0.0371

Table 3: Results for ablation studies on AASIST-L backbone.

System	ASVspoof 2021 evaluation set	
	EER	min t-DCF
AASIST-L	13.65	0.4574
RULE-AASIST-L	**12.91**	**0.4347**

Table 4: EER (%) and minimum t-DCF results for baseline and proposed model on the ASVspoof 2021 LA evaluation set.

enhancing the system's voice spoofing detection capabilities.

Notably, the proposed approach does not directly use the wav2vec 2.0 features as input to the spoofing detection model. Instead, it employs these representations to construct speech rules, which in turn modulate the output of high-level features F. This indirect usage of wav2vec 2.0 features contributes to the strong performance gains observed. As a result, the method opens promising directions for future research on using self-supervised representations to guide rule-based structures in voice spoofing detection.

4.4 Ablation Study

Table 3 presents the results of ablation experiments, in which individual components of the AASIST model are either removed or replaced. The results show a clear drop in performance when only the speech rule R is used as the high-level representation S. This performance degradation is attributed to the fact that R, while effective in modeling sequential consistency, lacks the rich acoustic information contained in the original high-level features F, making it insufficient on its own for effective spoofing detection. Similarly, replacing S directly with attention weights P results in an even more significant decline in performance. This suggests that attention weights alone, without the support of learned feature representations, are inadequate as standalone features.

Finally, we examine the effect of changing the source layer for feature extraction within the wav2vec 2.0 encoder. When the hidden states are extracted from layer 12 (the final layer) instead of layer 6 (an intermediate layer), a noticeable performance drop is observed. This can be explained by the representational nature of the final layer, which is optimized for ASR and tends to encode more abstract semantic features. While such features are useful for linguistic understanding, they often lack the lower-level acoustic cues that are critical for spoofing detection, thereby reducing detection effectiveness.

4.5 Cross-Corpus Evaluation

In this experiment, the ASVspoof 2021 LA evaluation set was further used to evaluate the cross-corpus performance of voice spoofing detection as shown in Table 4. It is evident that training solely on the ASVspoof 2019 training set and evaluating on the ASVspoof 2021 evaluation set leads to an increase in EER due to data mismatch. Nevertheless, the proposed RULE-AASIST-L model consistently outperforms the baseline AASIST-L, demonstrating that the wav2vec 2.0-based attention mechanism remains effective in improving the performance of voice spoofing detection in cross-corpus evaluations.

5 Conclusions

This work introduces RULE-AASIST-L, a rule-aware voice spoofing detection framework that utilizes attention-derived speech rules based on wav2vec 2.0 representations. By modeling the correlation between acoustic features and attention weights, the proposed method captures rule-based inconsistencies introduced by synthetic speech. Unlike previous methods that treat wav2vec 2.0 features as direct inputs, our approach exploits these representations to guide the learning of bona-fide speech patterns, thereby improving detection robustness. Experimental results on the ASVspoof 2019 LA dataset confirm the effectiveness of our method, with substantial performance gains over baseline systems. Ablation experiments further underscore the importance of rule modeling and the choice of representation layer, showing that intermediate-layer features (e.g., layer 6) retain richer acoustic cues than final-layer representations. In the future, this study opens new directions for integrating self-supervised learning and rule-based reasoning in the field of voice spoofing detection, and we plan to further investigate the possibility of utilizing the constructed speech rules during the inference stage without relying on wav2vec 2.0 features. One potential direction involves integrating alignment search and a flow-based module to generate approximated wav2vec 2.0 representations during inference, thereby eliminating the need for direct feature extraction from the original model.

Acknowledgments

This work was supported in part by the National Science and Technology Council (NSTC), Taiwan, under Grant No. NSTC 113-2222-E-218 -003 - MY2.

References

Y. Zhang, F. Jiang, and Z. Duan. 2021. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters, Volume 28*, pages 937–941.

J. Zhou, T. Hai, D. N. A. Jawawi, D. Wang, E. Ibeke, and C. Biamba. 2022. Voice spoofing countermeasure for voice replay attacks using deep learning. *Journal of Cloud Computing*, 11:51.

A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez. 2019. A light convolutional GRU-rnn deep feature extractor for ASV spoofing detection. In *Proceedings of INTERSPEECH*, pages 1068–1072.

J. Boyd, M. Fahim, and O. Olukoya. 2023. Voice spoofing detection for multiclass attack classification using deep learning. *Machine Learning With Applications*, 14:100503.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations," In *Proceedings of Neural Information Processing Systems (NeurIPS)*, pages 12449–12460.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing, Volume 29*, pages 3451–3460.

S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing, Volume 16, Number 6*, pages 1505–1518.

A. Bawitlung, S. K. Dash, and R. M. Pattanayak. 2025. Mizo Automatic Speech Recognition: Leveraging Wav2vec 2.0 and XLS-R for Enhanced Accuracy in Low-Resource Language Processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Z. Fan, M. Li, S. Zhou, and B. Xu. 2021. Exploring wav2vec 2.0 on speaker verification and language identification. In *Proceedings of INTERSPEECH*, pages 1509–1513.

B. Nasersharif and M. Namvarpour. 2024. Exploring the potential of Wav2vec 2.0 for speech emotion recognition using classifier combination and attention-based feature fusion. *The Journal of Supercomputing, Volume 80, Number 16*, pages 23667–23688.

H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *The Speaker and Language Recognition Workshop (Odyssey)*, pages 112–119.

H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans. 2021. End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*.

J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans. 2022. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 6367–6371.

M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee. 2019. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In *Proceedings of INTERSPEECH*, pages 1008–1012.

J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 47–54.

X. Wang and J. Yamagishi. 2021. A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection. In *Proceedings of INTERSPEECH*, pages 4259–4263.

Computational Approaches to Quantitative Analysis of Pause Duration in Taiwan Mandarin

I-Ping Wan*
Phonetics and Psycholinguistics Laboratory, Graduate Institute of Linguistics / Research Center for Mind, Brain, and Learning / Program in Teaching Chinese as a Second Language
National Chengchi University
ipwan@mail2.nccu.tw

Yu-Ju Lai
Phonetics and Psycholinguistics Laboratory
Graduate Institute of Linguistics
National Chengchi University
112555008@g.nccu.edu.tw

Pu Yu*
Phonetics and Psycholinguistics Laboratory
Graduate Institute of Linguistics
National Chengchi University
acadyupu@gmail.com

Abstract

This study presents a quantitative analysis of pause-duration patterns in a Mandarin spoken corpus to establish a baseline for prosodic and cognitive assessment. Drawing on cross-linguistic research, the distribution of pause patterns is viewed as reflecting multiple underlying factors. Longer pauses aligned with prosodic and syntactic boundaries indicate more deliberative and planned discourse rather than spontaneous speech. Such settings place higher demands on cognitive and articulatory planning, producing extended thinking time as speakers handle complex topics and specialized terminology. The spoken corpus was automatically processed and annotated using an in-house alignment and pause-tagging pipeline. Outlier detection with a 3.0×IQR threshold retained 35,474 tokens and removed extreme values exceeding 1,016 ms. Short and medium pauses remained stable across mean, median, and variability measures, while long pauses showed a moderate reduction (16,436 to 15,420 tokens), with mean duration decreasing from 535 to 426 ms and standard deviation sharply reduced from 786 to 169 ms, while the median stayed around 370–380 ms. These findings demonstrate that automatic cleaning primarily removed aberrant values while preserving linguistically meaningful long pauses. This baseline from non-impaired adult speakers underscores the need for corpus-specific frameworks and offers a reference point for cross-linguistic research on speech planning.

Keywords: Pause Duration, Speech Disfluency, Computational Approaches, Taiwan Mandarin, Spontaneous Speech, Quantitative Analysis

1 Introduction

Speech pauses and silences have been recognized as integral components of spoken interaction, reflecting cognitive processing, social norms and communicative strategy rather than mere absences of sound. Once viewed as interruptions, these pauses are now understood to serve important semantic, pragmatic and cognitive functions (Saville-Troike, 1985; Zuo, 2002). Socio-pragmatic studies have emphasized that silence, hesitations and pauses serve diverse interactional functions beyond turn-taking. For instance, Ephratt (2007) categorized pauses into four types including stillness, planned pauses, silencing and eloquent silence, highlighting their role as meaningful communicative acts. Similarly, Olaoye (2020) offered a typology of silence, including stillness, pauses, eloquent silence and judicial or commemorative silence, showing how these forms operate as communicative tools to express respect, humility, self-control, and conflict avoidance. By situating silence within sociolinguistic and pragmatic theory, these studies highlight silence as

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 116–123
November 20-22, 2025

a culturally and religiously embedded strategy with perlocutionary effects on interlocutors.

Building on this foundation, research on hesitation phenomena and filler use has offered valuable insights into language production processes. Modeling how speakers manage planning and execution through hesitations and disfluencies can inform both human–computer interaction and clinical applications. Grosjean and Collins (1979) long ago provided early evidence that speakers adjust breath and pause placement in read speech to match pre-planned syntactic structures, further linking silent pauses to deliberate production planning. A segment of silence exceeding 150 milliseconds in duration was classified as a speech pause (Maassen & Povel, 1984; Hammen et al., 1994). Moreover, a number of corpus studies of academic speech suggested that a relatively high proportion of long pauses align closely with prosodic or syntactic boundaries. This pattern is reminiscent of formal or highly prepared speech genres such as reading aloud and political speeches (Duez, 1982; Grosjean & Collins, 1979; Ferreira, 1993). Also, Ferreira (1993) argued that prosodic planning, rather than purely syntactic parsing, governs pause insertion, especially at sentence ends, a view echoed by Krivokapić et al. (2020), who treated grammatical pauses as anticipatory prosodic boundary events. Zellner (1994) also emphasized the close link between pauses, prosody and information packaging. In a prepared speech, pauses are more structurally aligned and semantically functional. Other work also confirmed style-sensitive variation where Gustafson-Capkova et al. (2001) observed systematic differences in pause placement, frequency and duration across spontaneous dialogue, amateur reading and professional broadcasting. In the latter, pauses were shorter, less frequent and more tightly aligned with syntactic boundaries, consistent with higher planning and rhetorical control.

In terms of pause patterning, Campione and Véronis (2002) compared pause patterns in read vs. spontaneous speech across five languages and found that read speech exhibited a more regular bimodal distribution of short and medium-length pauses. In contrast, spontaneous speech introduced a third mode, which showed rather long pauses (often >1000 ms), typically associated with hesitation, lexical search, or real-time syntactic planning. This suggests that formal and pre-planned speech tends to contain longer structurally aligned pauses, while extremely long pauses are characteristic of high planning load in spontaneous dialogue. In their studies, a methodological caveat emerges when setting the lower boundary of "long pause" at >250 ms: such a threshold may conflate two functionally distinct phenomena including boundary-aligned silences in formal registers and hesitation-induced delays in spontaneous speech. It is critical to distinguish these planned boundary-aligned pauses from extremely long pauses that more likely reflect spontaneous cognitive planning difficulties (Campione & Véronis, 2002). Šturm (2023) further compared news reading with poetry reading and demonstrated that pause patterns are shaped not only by genre but also by the underlying text structure (explicit vs. implicit cues), highlighting how increasing planning demands and formality elevate discourse-based pause control.

Computational and empirical approaches have sought to model disfluencies and pause phenomena in large-scale speech data. Aijmer (2011) and Crible (2017) utilized prosodic features as cues to indicate the presence of prosodic markers. Betz et al. (2020) investigated the form, function and modeling of disfluencies, especially hesitations, in human speech and their integration into spoken dialogue systems, providing empirical data on the frequency, distribution and acoustic characteristics of silent and filled pauses. Similarly, Wan and Allassonnière-Tang (2021) present a connectionist model of Mandarin speech production to examine how word frequency and position within an utterance influence the occurrence of speech errors, using corpus-based data and computational simulations. Zhang (2024) further applied quantitative methods to spontaneous speech corpora to uncover sociolinguistic variation linked to speech planning. These findings suggested that features such as pause duration, frequency and distribution can support speaker-state detection and automatic speech processing, extending the relevance of pause research beyond linguistics into computational and even forensic applications.

Clinical and cognitive research has increasingly begun to explore pauses and silences as sensitive markers of neurological and cognitive status. Imre et al. (2022) analyzed silent pauses, hesitations and irrelevant utterances in phonemic and semantic fluency tasks, demonstrating that silence-related parameters such as the length of pauses can effectively differentiate between individuals with mild cognitive impairment and healthy controls. In a complementary study, Sluis et al. (2020) presented an automated approach to analyzing pausing behavior in the speech of people with dementia using the Calpy open-source speech processing toolkit. They found progressive increases in pause duration and proportion of silence across groups, alongside a rise in very long pauses (≥2000 ms) and decreases in total speech duration and mean phrase length, demonstrating that automated pause detection can effectively capture speech disfluencies associated with dementia and support future diagnostic and communication research.

Therefore, these strands of research indicate that pauses and silences are multi-layered phenomena bridging sociocultural, cognitive, and computational domains. However, despite substantial advances, most of this work has been conducted on English or other major European languages, and there remains a paucity of comparable studies in Mandarin. Chen et al.(2022) further examined how discourse functions are reflected through phonological or acoustic features. However, there is still a lack of integrated corpora that combine detailed pause-duration measurements from healthy speakers with the methodological rigor necessary for later comparison to clinical populations. Existing studies either focus on the qualitative or typological aspects of silence, or they apply automated methods primarily to clinical or task-based data without establishing a robust baseline from non-impaired speech in naturalistic settings. Therefore, this study aims to fill this gap by constructing a quantitative corpus-based resource of pause duration in Taiwan Mandarin, providing a robust baseline of silent and filled pauses in naturalistic speech. This corpus not only enables direct comparison with existing English-language studies but also lays the groundwork for future research on aging and clinical populations.

In this paper, we address this gap by constructing a speech-pause corpus that provides high-quality and time-aligned pause data from non-impaired speakers. This corpus is designed to support cross-sectional and longitudinal analyses of pause duration and distribution. By combining socio-pragmatic insights with computational modeling and corpus-based methods, our approach aims to advance both theoretical understanding and practical applications of pause analysis in naturalistic speech. Ultimately, we envision that this resource can be extended to high-risk and aging populations, enabling comparative research on pause behavior as an indicator of cognitive and communicative change in the near future. Questions to be investigated include the following:

1. How can a dedicated speech-pause corpus of non-impaired speakers be designed and annotated to capture detailed pause-duration information across spontaneous speech?
2. To what extent do pause-related parameters, such as number of pauses, average pause length or distribution, provide a reliable baseline for future comparisons with aging and clinical populations?
3. How can insights from socio-pragmatic studies of silence and computational modeling of disfluencies be integrated to improve the automatic detection and classification of pause phenomena?
4. In what ways can such a corpus support cross-linguistic or cross-task analyses, enabling the identification of sociolinguistic variation and potential early markers of cognitive decline?

2 Methodology

A subset of the corpus, totaling 16 hours, 8 minutes, and 2 seconds, drawn from a larger 202-hour multimodal Mandarin speech database, was automatically annotated using Praat (Boersma & Weenink, 2023–2025) for fine-grained analysis of features such as fillers and silent pause-related phenomena. This section outlines the participants, data collection procedures, annotation schema and analysis methods used in the study.

All participants were native speakers of Taiwan Mandarin (N = 4; 1 male, 3 females; age range = 23–25 years, M = 24.2, SD = 0.7). Although the corpus size used here is relatively limited, it was intentionally designed as a controlled case study focusing on young adult speakers with comparable linguistic and cognitive profiles. The goal of this study is not large-scale modeling, but to provide a proof-of-concept analysis demonstrating how automatic annotation can reveal pause and filler patterns in naturalistic speech.

Recordings were made in controlled environments using high-quality audio equipment. The primary content comprises graduate-level classroom settings, including instructor lectures and interactive seminar-style discussions between instructors and students. Notably, over 97% of the annotated utterances showed no statistical outliers in pause duration, indicating a high degree of internal consistency and reliability in the dataset. The combination of academic lectures, seminar discussions, free conversations, and short cognitive-linguistic exercises ensures a rich distribution of spontaneous speech, encompassing a wide range of pause types and speech planning demands.

Drawn partially from graduate classroom discussions, the corpus represents a semi-spontaneous academic register rather than a fully unplanned conversation. However, the speakers produced their utterances without any prepared script or reading material, and the recordings capture natural pauses, hesitations and fillers characteristic of spontaneous speech production. This makes the data appropriate for a case study of cognitive and prosodic pause behaviors in controlled academic discourse, which complements findings from more casual conversational corpora.

Regarding the nature of our speech data, we agree that some portions of the corpus (e.g., classroom lectures) may reflect a more deliberative and planned register. However, these data were chosen because they still involve spontaneous verbal responses, turn-taking, and hesitations typical of natural speech in academic contexts.

We employed an in-house automatic phonetic alignment pipeline developed and refined over several years in the laboratory, rather than relying on open-source tools.[1] This system, combined with manual verification, allows for highly accurate segmentation and annotation. Pauses are operationalized as segments of silence or silent pauses detected by our automated tagging procedure. Each pause instance is annotated with start time, end time, duration, and position relative to syntactic boundaries. From these annotations, we extracted the number of pauses, mean pause duration and distributional patterns from spontaneous speech. Metadata included various speakers, speech type and speech rate. These measures in the future hope to provide a normative baseline for future comparisons with aging or clinical populations. Quantitative analyses include descriptive statistics to identify pause-duration profiles.

Pause duration was identified by detecting segments of silence in the acoustic waveform. According to Maassen & Povel (1984) and Hammen et al. (1994), the data were categorized based on two duration thresholds, which were 150 milliseconds and 250 milliseconds, resulting in three distinct groups: pauses shorter than 150 ms, pauses between 150 and 250 ms, and pauses exceeding 250 ms.

3 Data Analysis

We first tested whether data cleaning materially altered the distribution of pause categories, as shown in Figure 1.

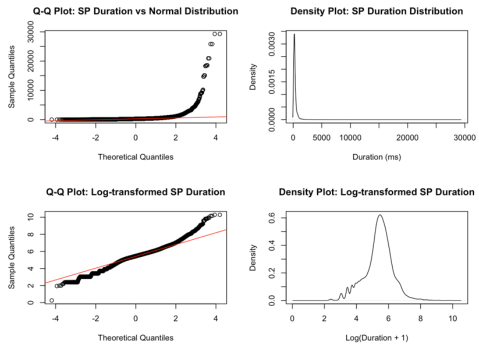


Figure 1: Distributional characteristics of speech pause duration in raw and log-transformed scales .

Pause durations showed a pronounced right-skewed, non-normal distribution (n = 36,490), with a peak around 230 ms and a long upper tail. Long pauses were disproportionately affected by outlier removal: 1,016 long pauses were excluded, reducing the mean from 535 to 426 ms (−20%) and compressing the standard deviation (786 to 169 ms). These changes primarily truncated extreme hesitations rather than altering the median (380 to 370 ms), suggesting that the core distribution of boundary-aligned pauses remained stable. The cleaned corpus therefore reflects a clearer

[1] The speech data were processed using a self-supervised in-house phonetic alignment pipeline developed with Praat scripting and custom Python routines, rather than relying on forced-alignment toolkits. The system performs automatic segmentation, boundary detection and iterative self-correction through acoustic feature learning, allowing cross-linguistic adaptability (see Wan et al., 2024, for how Thai preschoolers learn Mandarin). This study, however, does not address prosody or intonation, as its primary focus lies in the analysis of pause and hesitation phenomena within spontaneous speech.

distinction between short/medium pauses, which often aligned with prosodic or syntactic boundaries, and very long pauses, which tend to index planning or hesitation in spontaneous speech (cf. Campione & Véronis, 2002). Because pause durations are non-normally distributed, non-parametric methods and median/IQR statistics are used. This approach preserves linguistically meaningful contrasts between routine boundary pauses and hesitation-driven silences, while minimizing the influence of outliers.

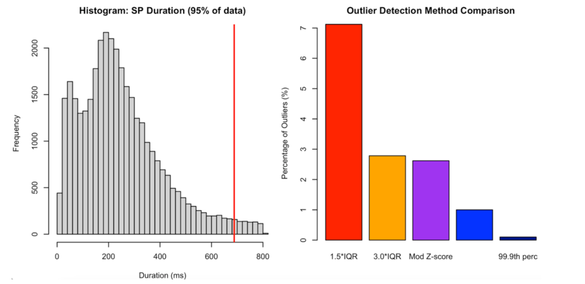


Figure 2: Detection of outliers in pause durations

In Figure 2, outlier detection analysis revealed substantial upper-tail extremes requiring data cleaning prior to modeling. Using the 3.0×IQR method (threshold ≈1,016 ms) alongside a modified Z-score approach (|z| > 3.5), we identified approximately 3% of pauses as outliers. In contrast, the standard 1.5×IQR criterion flagged over 7% of pauses, which was deemed overly restrictive for preserving natural speech variability. Frequency analysis of the central 95% of data showed a right-skewed distribution peaking around 200–300 ms, consistent with known pause distributions in formal and semi-formal speech. By selecting the 3.0×IQR criterion, we retained linguistically meaningful long pauses while trimming only extreme hesitation events, resulting in a final dataset of 35,474 observations capped at about one second. This procedure preserves the contrast between short/medium pauses—often aligned with prosodic or syntactic boundaries—and very long pauses, which tend to index planning or hesitation in spontaneous speech (Campione & Véronis, 2002).

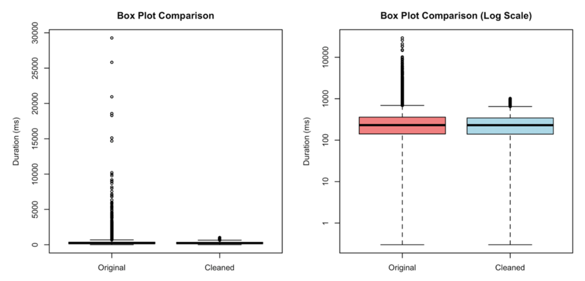


Figure 3: Impact of outlier removal on speech pause duration distributions.

As shown in Figure 3, outlier removal using the 3.0×IQR threshold produced minimal impact on the central distribution while substantially reducing extreme variability. Median pause duration remained essentially unchanged (≈230 ms), while the interquartile range decreased moderately, indicating that core pause behavior was preserved. The most pronounced effect was the elimination of extreme upper outliers without distorting the underlying distribution. Short pauses (<150 ms; n=9,767) and medium pauses (150–250 ms; n=10,287) were unaffected by data cleaning, retaining virtually identical means, medians, and standard deviations. In contrast, long pauses (>250 ms) showed the largest adjustment (n reduced from 16,436 to 15,420), with mean duration decreasing from about 535 to 426 ms and standard deviation sharply reduced, while the median shifted only slightly (380→370 ms). This selective effect confirms that the procedure primarily targeted aberrant values in the upper tail while preserving linguistically meaningful pause patterns. Short and medium pauses continue to represent routine boundary-aligned silences, whereas the cleaned long-pause category better reflects legitimate planning-related hesitations rather than measurement noise, aligning with established pause typologies (Campione & Véronis, 2002).

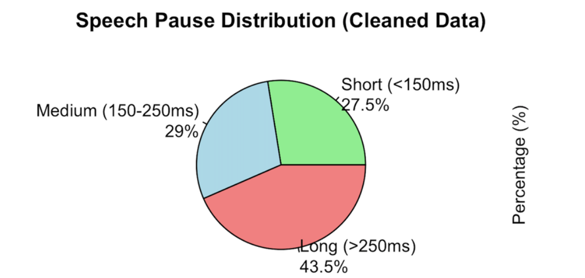


Figure 4: Speech pause distribution by duration category following data cleaning.

The present analysis revealed a distinctive pause distribution pattern that deviates substantially from typical conversational speech norms reported in the literature (Figure 4). The observed distribution—short pauses (<150ms): 27.5%, medium pauses (150-250ms): 29.0%, long pauses (>250ms): 43.5%—contrasts markedly with expected ranges where short pauses typically comprise 40-60% and long pauses 10-30% of total pause events. This inverted pattern, characterized by a predominance of long pauses and relative scarcity of brief hesitations, suggests speech production involving heightened cognitive processing demands rather than spontaneous discourse.

Several factors may account for this distributional profile. The elevated proportion of long pauses likely reflects deliberative speech planning processes, indicating that speakers engaged in more cognitively demanding language production requiring additional processing time for lexical access, syntactic formulation, or discourse organization. The reduced frequency of micropauses and brief hesitations suggests less spontaneous, more controlled speech output characteristic of formal register or task-specific contexts. This pattern is consistent with speech elicited in academic interviews, formal presentations, or complex narrative tasks where speakers prioritize accuracy and coherence over fluency.

The linguistic implications extend beyond simple temporal measurements to suggest fundamental differences in speech production mechanisms. The predominance of longer articulatory timing intervals may indicate enhanced monitoring processes, increased attention to phonetic precision, or elevated cognitive load associated with L2 speech production or specialized discourse domains. These findings underscore the importance of considering contextual factors when interpreting pause patterns and highlight the need for corpus-specific normative data in prosodic boundary analysis.

Within-category frequency analysis revealed distinct distributional characteristics across pause types in the cleaned dataset (n = 35,474), as shown in Figure 5. Short pauses (<150ms) exhibited a right-skewed distribution with modal frequency around 80-90 ms and high consistency between median (80.00 ms) and mean (83.32 ms), indicating minimal internal variability. Medium pauses (150-250ms) demonstrated the most symmetric distribution with peak frequency at 200 ms and perfect convergence of median and mean values (200.00 ms), reflecting highly standardized phrase boundary timing. Long pauses (>250ms) showed pronounced right skew with median (370.00 ms) substantially lower than mean (426.24 ms), indicating considerable internal heterogeneity despite outlier removal. The long pause category maintained an extended upper tail reaching the 1,016 ms threshold, suggesting that even within linguistically valid boundaries, substantial variation exists in processing-related articulatory timing intervals.

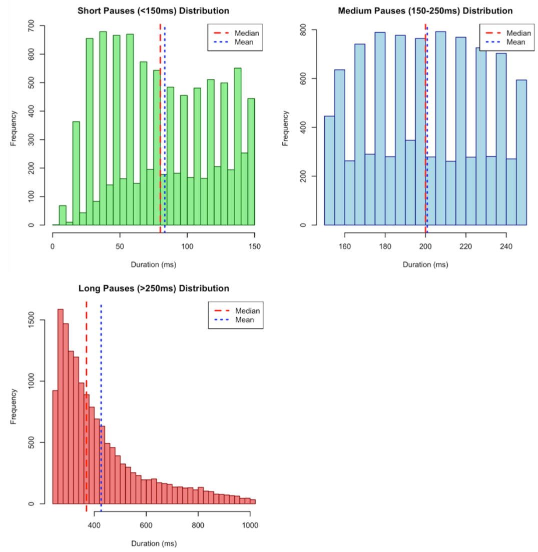


Figure 5: Frequency distributions of speech pause duration by category following outlier removal.

This figure effectively demonstrates that the three-category classification captures fundamentally different pause phenomena, with each category showing distinct statistical properties that justify separate analytical treatment.

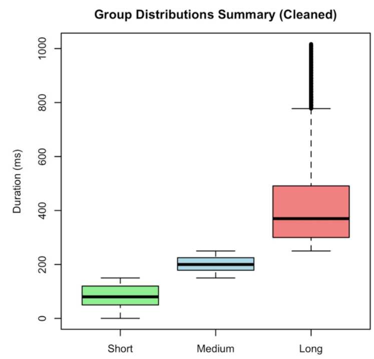


Figure 6: Comparative box plot distributions of pause duration categories in cleaned dataset.

Box plot comparison across pause categories confirmed distinct non-overlapping duration ranges with varying internal distributions following data cleaning. Short pauses demonstrated tight clustering with median at 80.00 ms, narrow interquartile range (51.14-110.00 ms), and minimal outliers, indicating highly consistent micropause timing. Medium pauses exhibited the most compact distribution with median at 200.00 ms and symmetrical quartile spacing (174.36-

225.64 ms), reflecting standardized phrase boundary durations. Long pauses showed the greatest variability despite outlier removal, with median at 370.00 ms, expanded interquartile range (290.91-481.09 ms), and extended upper whiskers reaching the 1,016 ms threshold. The clear separation between categories validates the literature-based classification scheme, while the progressive increase in variability from short to long pauses reflects the transition from automated articulatory timing to cognitively-mediated processing intervals.

This box plot effectively summarizes the key finding that the three categories represent genuinely distinct pause phenomena with different underlying timing mechanisms.

For word-level pauses labeled sp, cleaning produced a small but reliable shift in category composition: the proportion of long pauses decreased by 1.58 percentage points (from 45.04% to 43.47%), with corresponding increases in short (+0.77 pp, from 26.79% to 27.56%) and medium (+0.81 pp, from 28.17% to 28.98%). The association between dataset (Original vs. Cleaned) and category was significant, χ^2 (2) = 18.10, p = 1.17×10^{-4}, Cramér's V = 0.016 (small effect). Because the cleaned set is a subset of the raw set, this test quantifies a composition shift rather than independence.

The significant change is expected given the rule that removes extremely long pauses; the effect size is small (V≈0.016), indicating that cleaning mainly trims the right tail without materially altering central tendencies. Substantively, inferences about typical pause behavior should remain stable, while metrics sensitive to heavy tails (e.g., variance, mean) become less influenced by outliers. For rigor, if token-level retention flags are available, a paired/marginal-homogeneity test can confirm the finding; additionally, a sensitivity analysis across alternative cutoffs (e.g., 800–1,200 ms) can demonstrate robustness.

4 Conclusion

Based on cross-linguistic research findings, the distinctive distributional pattern observed in this Chinese corpus likely reflects several underlying factors. The speech production characteristics suggest more deliberative and planned discourse, potentially originating from formal or academic contexts with reduced spontaneous rapid speech. The cognitive processing patterns indicate increased demands for articulatory planning, with language production involving extended thinking time that may reflect topic-specific complexity or cognitive load.

The data collection context provides an interpretive framework for these findings. The corpus appears to derive from structured interactions such as interviews, presentations, or academic discussions, where speakers engage with specialized content requiring careful formulation. Notably, the speakers may represent non-native Chinese users presenting advanced academic material, a context that inherently promotes more cautious speech production with extended processing intervals. This linguistic environment naturally facilitates longer pause durations as speakers navigate complex conceptual material while managing potential language proficiency constraints.

These findings underscore the importance of contextual factors in prosodic boundary analysis and highlight how discourse demands, speaker characteristics, and communicative settings interact to shape temporal speech patterns. The results provide valuable baseline data for understanding pause distributions in formal Chinese academic discourse and demonstrate the necessity of corpus-specific normative frameworks for cross-linguistic prosodic research.

Acknowledgments

We appreciate the valuable and constructive comments from two anonymous reviewers. Our deepest appreciation goes to Dr. Chain-wu Lee for his continuous cutting-edge high-tech programming support in constructing all the corpora in the Phonetics and Psycholinguistics lab at National Chengchi University. All remaining errors in the analysis of interpretation are our own. This research was supported in part by the Ministry of Education, Taiwan, under the International Integrated Collaboration Project for the University Alliance in the Czech Republic (ICU) & University Academic Alliance in Taiwan (UAAT) to the first author (113F113085).

References

Aijmer, Karin. (2011). Well I'm not sure I think… The use of well by non-native speakers. *International Journal of Corpus Linguistics*, 16, 231–254.

Betz, Simon, et al. (2020). Hesitation Processing Analysis Using Continuous Mouse-Tracking and Gamification. In *Elektronische Sprachsignalverarbeitung 2020. Tagungsband der 31. Konferenz.* (Vol. 95).

Boersma, Paul & Weenink, D. (2023-2025). *Praat: Doing phonetics by computer* [Computer program] (Version 6.4.27). Retrieved January 27, 2025, from http://www.praat.org/

Campione, Estelle, & Jean Véronis. (2002). A large-scale multilingual study of silent pause duration. In *Speech prosody* (Vol. 2002, pp. 199-202).

Chen, Pin-Er, et al. (2022). Analyzing discourse functions with acoustic features and phone embeddings: Non-lexical items in Taiwan Mandarin. *International Journal of Computational Linguistics & Chinese Language Processing*, 27(2).

Crible, L. (2017). Discourse markers and (dis) fluency across registers. *A Contrastive Usage-Based.*

Duez, Danielle. (1982). Silent and non-silent pauses in three speech styles. *Language and speech*, 25(1), 11-28.

Ephratt, Michal. (2007). On silence—introduction. *In Silence in Culture and in Interpersonal Relations* (pp. 7–25). Resling: Tel Aviv.

Ferreira, Fernanda. (1993). Creation of prosody during sentence production. *Psychological review*, 100(2), 233.

Grosjean, François, & Maryann Collins. (1979). Breathing, pausing, and reading. *Phonetica*, 36, 98–114.

Gustafson-Capkova, Sofia, & Beata Megyesi. (2001). A comparative study of pauses in dialogues and read speech. In *INTERSPEECH* (pp. 931-934).

Hammen, Vicki L., Kathryn M. Yorkston, and Fred D. Minifie. (1994). The effect of temporal alterations on speech intelligibility in Parkinsonian dysarthria. *Journal of Speech and Hearing Research*, 37(2), 244–253. https://doi.org/10.1044/jshr.3702.244

Imre, Nóra, et al. (2022). Temporal speech parameters indicate early cognitive decline in elderly patients with type 2 diabetes mellitus. *Alzheimer Disease & Associated Disorders*, 36(2), 148-155.

Krivokapić, Jelena, Will Styler, & Benjamin Parrell. (2020). Pause postures: The relationship between articulation and cognitive processes during pauses. *Journal of Phonetics*, 79, 100953.

Maassen, Ben, & Dirk-Jan Povel. (1984). The effect of correcting temporal structure on the intelligibility of the deaf. *Speech Communication*, 3(2), 123–135. https://doi.org/10.1016/0167-6393(84)90034-7

Olaoye, Anthony Ayodele. (2020). A socio-pragmatic analysis of silence in communication: An ethnographic review. *Veritas Journal of Humanities*, 2(1), 108–113.

Saville-Troike, M. (1985). The place of silence in an integrated theory of communication. In D. Tannen & M. Saville-Troike (Eds.), *Perspectives on Silence* (pp. 3–18). Norwood, NJ: Ablex.

Sluis, Rachel A., et al. (2020). An automated approach to examining pausing in the speech of people with dementia. *American Journal of Alzheimer's Disease & Other Dementias*, 35, 1533317520939773.

Šturm, Pavel, & Jan Volín. (2023). Occurrence and duration of pauses in relation to speech tempo and structural organization in two speech genres. *Languages*, 8(1), 23.

Wan, I-Ping & Allassonnière-Tang, Marc. (2021). The effect of word frequency and position-in-utterance in Mandarin speech errors: A connectionist model of speech production. *The Post-conference proceedings of CLSW2020 in the LNAI series*. Springer Singapore. 491-500.

Wan, I-Ping and Marc Allassonnière-Tang & Pu Yu. (2024). Early segmental production in Thai preschool children learning Mandarin. *International Journal of Asian Language Processing, 34*(2), 2450005.

Zhang, Jing. (2024). Variational pragmatics in Chinese discourse markers zhege and nage: The influence of region and gender. *Journal of Pragmatics,* 230, 76–88.

Zuo, Yan. (2002). The golden silence: A pragmatic study on silence in dyadic English conversation. *München: Lincom Europa*.

基於隱藏式馬可夫模型的中文熟語自動糾錯新方法
A Novel Chinese-Idiom Automatic Error Correction Method Based on the Hidden Markov Model

張榕彬 Rongbin Zhang[1*], 桂安露 Anlu Gui[1], 曹鵬 Peng Cao[1]
吳凌峰 Lingfeng Wu[1], 黃鳳 Feng Huang[1], 李家暉 Jiahui Li[2]
[1]Department of Electronic Engineering, Shantou University, Shantou, China
[2]College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou, China
{rbzhang, 24algui, 22pcao, 22lfwu, 22fhuang1}@stu.edu.cn
qsx20240853@student.fjnu.edu.cn

摘要

在日常學習和使用中，受各種錯別字和光學字元辨識誤差等影響，中文熟語的拼寫經常出現錯誤。實現熟語自動糾錯是中文自然語言處理的重要任務之一，有助於提升中文文本品質和漢語學習效果。現有的編輯距離方法和自定義詞典方法，都存在糾錯能力受限、計算效率較低、靈活性不足等問題。鑒於此，本文提出一種基於隱藏式馬可夫模型（hidden Markov model，HMM）的中文熟語自動糾錯方法，將熟語錯誤的產生過程用 HMM 模型來建模，把熟語糾錯問題轉化爲錯誤熟語與合法熟語之間的匹配問題。通過構建合法熟語表和漢字混淆集，開發一個熟語自動糾錯原型系統，並完成性能測試。實驗表明，與現有方法相比，本方法模型簡單、參數少、計算複雜度低，具備更強的糾錯能力和參數健壯性，能夠更靈活地糾正多樣化類型的熟語錯誤，具有較高的應用潛在價值。

Abstract

Spelling errors in Chinese idioms frequently occur due to various types of misspellings and optical character recognition errors in daily learning and usage. Achieving automatic error correction for Chinese idioms is one of the important natural language processing tasks, as it helps improve the quality of Chinese texts as well as language learning. Existing methods, such as edit distance and custom dictionary approaches, suffer from limited error correction capability, low computational efficiency, and weak flexibility. To address these limitations, this paper proposes a novel automatic error correction method for Chinese idioms based on the hidden Markov model (HMM). Specifically, the generation process of idiom spelling errors is modeled using an HMM, transforming the idiom correction problem into a matching task between erroneous idioms and legitimate idioms. By constructing a legitimate idiom table and a Chinese character confusion set, a prototype system for idiom correction was developed, and performance testing was completed. Experiment results demonstrate that the proposed model is simpler with fewer parameters and has lower computational complexity while exhibiting stronger error correction capability and parameter robustness as compared to existing methods. It can more flexibly correct diverse types of idiom errors, showing high potential application value.

關鍵字：中文熟語、自動糾錯、隱藏式馬可夫模型

Keywords: Chinese idiom, Automatic error correction, Hidden Markov model

1 緒論

中文是世界上最古老的文字之一，其歷史悠久，沉澱了豐富的詞彙和熟語。其中，熟語（idiom）是指人們從日常生活經驗中總結出來的短語或短句，通常言簡意賅、含義深刻，例如成語、慣用語、歇後語、諺語、格言、詩詞名句等。然而，由於熟語結構特殊、用法靈活，在日常學習和使用時，容易出現拼寫錯誤。常見的熟語拼寫錯誤來源，有近形字、同音字、近音字、光學字元辨識錯誤、鍵盤誤操作，最終表現爲漢字重複、缺失和誤用等。研究實現中文熟語自動糾錯，是自然語言處理領域的問題之一，對於檢驗中文教學效果、提升中文文本品質等具有重要意義。

本文以隨機過程、時間序列和機率統計的視角，重新思考熟語拼寫錯誤的產生原因和應對方法，開發一種靈活糾正多樣化錯誤類型的中文熟語自動糾錯輕量型方案。我們的靈感來自於隱藏式馬可夫模型（hidden Markov model，HMM）在語音辨識等領域的應用（Rabiner, 1989）。在針對孤立詞（isolated words）的語音辨識中，HMM 模型用於建模語音信號的序

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 124–132
November 20-22, 2025 ©2025 Association for Computational Linguistics

Error	Correction	Transformation Correct Letter	Error Letter	Position (Letter #)	Type
acress	actress	t	—	2	deletion
acress	cress	—	a	0	insertion
acress	caress	ca	ac	0	transposition
acress	access	c	r	2	substitution
acress	across	o	e	3	substitution
acress	acres	—	s	5	insertion
acress	acres	—	s	4	insertion

Figure 1: 與錯誤詞 acress 編輯距離爲 1 的合法候選詞示例。截圖自 (Jurafsky and Martin, 2024) 的 Appendix B

列特徵，利用模型的狀態轉移機率和觀測機率等參數，經運算後對說話者表達的詞語或句子進行辨識。Lin et al. (2012) 曾利用 HMM 模型實現一個英文單詞拼寫檢查器的簡單原型系統，其思想與 HMM 模型在語音辨識中的應用有異曲同工之妙。HMM 模型在時間序列分析中所展現的靈活性和健壯性，爲本方法提供了寶貴靈感和理論基礎。

現有的熟語拼寫自動糾錯方法，主要有兩種。第一種是基於編輯距離（edit distance，也稱 Damerau-Levenshtein 距離）(Hodge and Austin, 2003; Wang et al., 2014) 的方法，常用於英文單詞拼寫檢查 (Jurafsky and Martin, 2024; Norvig, 2016; Revathi et al., 2023)，隨後也用於俄語 (Varlamova et al., 2023)、孟加拉語 (Khairul Islam et al., 2019)、緬甸語 (Mon et al., 2021)、印度尼西亞語 (Soleh and Purwarianti, 2011)、印地語 (Jain and Jain, 2014) 等單詞糾錯。該方法需要根據給定的錯誤詞，產生具有編輯距離的可能詞作爲候選集，進而在合法詞彙表中進行查詢和匹配。其中，編輯距離是指字元插入（insertion）、字元刪除（deletion）、字元替換（substitution）、相鄰字元換位元（transposition）等基本操作的次數，如圖1所示。然而，由於優選集的大小隨著編輯距離的增加呈現指數級增長，該方法通常只能處理具有較小編輯距離的拼寫錯誤，其糾錯能力受到極大的限制。

第二種是自定義詞典方法，即首先在詞典中列出常見的錯誤詞及其對應的正確詞，進而將錯誤詞與詞典中的樣本逐一進行比對。典型案例是微軟公司 Word 辦公軟體自帶拼寫檢查器 (Microsoft-Corporation, 2025) 的成語糾錯功能，如圖2所示。然而，該方法只能處理已被收錄的特定拼寫錯誤，因而靈活性低，糾錯能力有限。例如，根據圖2的詞典，該方法能夠將「冰上巴蕾」正確地糾正爲「冰上芭蕾」，但無法對「冰上巴雷」進行檢錯和糾錯，因爲詞典並未收錄「冰上巴雷」這一錯誤詞。

針對現有方法在糾錯能力、計算效率和靈活性等方面的局限，我們使用 HMM 機率模型，

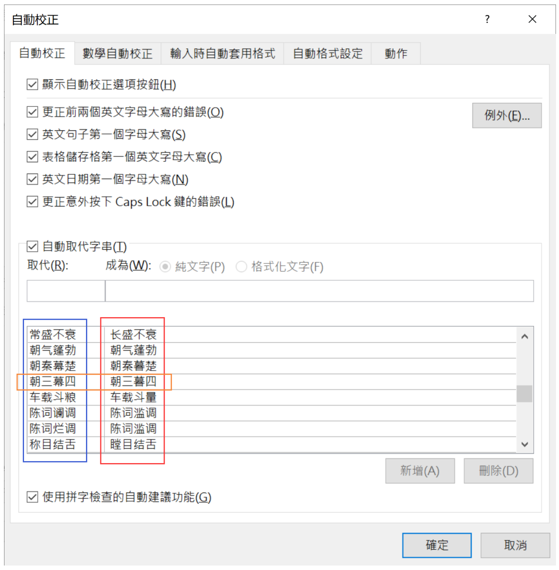


Figure 2: 微軟公司 Word 辦公軟體拼寫檢查器只能糾正已被收錄的錯誤成語。截圖自：Microsoft Word -> 檔案 -> 選項 -> 校訂 -> 自動校正選項 -> 自動校正

實現一種中文熟語自動糾錯輕量級方法，以期對於多樣化類型的熟語使用錯誤都能夠靈活地自動糾正。本論文的貢獻如下：

1) 以隨機過程、時間序列和機率統計等新視角，重新看待中文熟語糾錯問題，使用 HMM 模型進行問題建模和求解。

2) 在問題建模中，對於 HMM 模型的轉移機率和觀測機率等關鍵參數，都巧妙地賦予物理意義，使得所提出的自動糾錯演算法具有很強的合理性和可解釋性。

3) 與基於編輯距離的傳統方法相比，本方法的靈活性更強，可糾正不同類型的中文熟語使用錯誤，而不局限於簡單的少量字元替換、刪除或插入錯誤。

4) 與基於自定義詞典的傳統方法相比，本方法的糾錯能力更強，而不局限於詞典中預先收錄的特定錯誤詞。

5) 本文所提出演算法模型簡單、參數少、計算效率高、可擴展性強，可根據實際需求即時更新合法熟語表，可靈活地應用於成語、慣用語、歇後語、諺語、格言、詩詞名句等的自動糾錯任務。

2 提出方法

2.1 HMM 基礎

HMM 模型最初由 Leonard E. Baum 等提出，後來 Lawrence R. Rabiner 等進行深入研究 (Rabiner, 1989)。該模型是一種雙重隨機過程，表現在其狀態轉換過程是隱藏的、無法直接觀測的，而可觀測事件的隨機過程是隱藏

狀態轉換過程的隨機函數 (Zong, 2024)。

習慣上，使用五元組 $\lambda = (N, M, \mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ 來表示一個離散平穩 HMM 模型，參數包括：

1) 隱藏狀態的個數 N；

2) 觀測狀態的個數 M；

3) 轉移機率矩陣（transition matrix）$\mathbf{A}$。該矩陣是 $N \times N$ 方陣，其第 i 行第 j 列元素 $[\mathbf{A}]_{i,j}$ 表示任意時刻 t 由隱藏狀態 s_i 轉移爲下一時刻 $t+1$ 的隱藏狀態 s_j 的機率；

4) 觀測機率矩陣（也稱混淆機率矩陣、發射機率矩陣，observation/confusion/emission matrix）$\mathbf{B}$。該矩陣有 N 行 M 列，其第 j 行第 k 列元素 $[\mathbf{B}]_{j,k}$ 表示任意時刻 t 由隱藏狀態 s_j 產生觀測狀態 v_k 的機率；

5) 初始機率向量（initial probability vector）$\boldsymbol{\pi}$。該向量是長度爲 N 的行向量（row vector），其第 i 個元素 $[\boldsymbol{\pi}]_{1,i}$ 表示在初始時刻 $t=1$ 系統處於隱藏狀態 s_i 的機率。

HMM 模型涉及評估（evaluation）、解碼（decoding）、訓練（training）等三個基本問題 (Rabiner, 1989; Zong, 2024; Stamp, 2021)。其中，評估問題描述爲：給定 HMM 模型 $\lambda = (N, M, \mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ 和觀測狀態序列 $\mathbf{o} = o_1 o_2 \cdots o_T$，計算在該模型下產生序列 $\mathbf{o}$ 的機率 $\Pr\{\mathbf{o}|\lambda\}$。該問題可使用前向演算法（Forward Algorithm）進行求解。本文所提出的方法將使用評估問題來建模並求解。

2.2 熟語自動糾錯的 HMM 建模

本方法的切入點在於爲合法熟語表中的所有熟語分別建立各自的 HMM 模型，而關鍵在於如何合理設置每個 HMM 模型的參數。

我們將待糾錯的錯誤熟語 $\mathbf{w}$ 視爲觀測狀態序列，例如含有拼寫錯誤的「冰上巴雷」。假設所使用的合法熟語表共有 N_{idiom} 個熟語，例如 THUOCL 詞庫提供的合法成語表 (Han et al., 2016) 共有 $N_{\text{idiom}} = 8519$ 個成語。記第 n 個合法熟語爲 $\mathbf{c}_n\,(n = 1, 2, \cdots, N_{\text{idiom}})$，其對應的 HMM 模型爲

$$\lambda_n = (N_n, M_n, \mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n). \tag{1}$$

模型 λ_n 產生觀測狀態序列 $\mathbf{w}$ 的機率爲 $\Pr\{\mathbf{w}|\lambda_n\}$。此時，一種合理的策略是取

$$\mathbf{c}_{\text{opt}} = \arg\max_n \Pr\{\mathbf{w}|\lambda_n\} \tag{2}$$

作爲最佳糾錯建議。式 (2) 所使用的最優準則，可以看成是錯誤熟語與所有合法熟語之間的最佳匹配。[1]

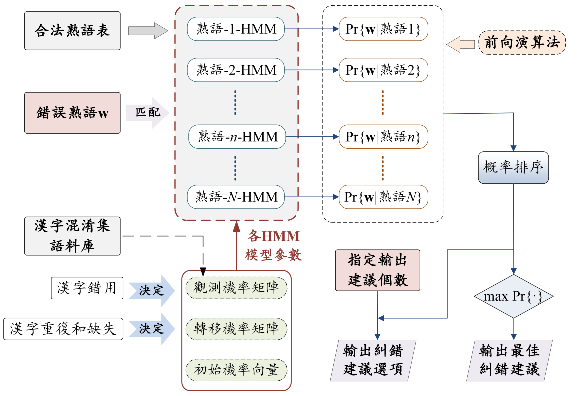


Figure 3: 基於 HMM 模型的中文熟語自動糾錯方法實施方案

餘下的問題是如何爲每個合法熟語確定 HMM 模型參數 $\{N_n, M_n, \mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n\,\forall n\}$。爲此，需要重新思考熟語使用錯誤的產生機理，進而確定建模方法。

首先，我們將隱藏狀態個數 N_n 設置爲合法熟語 $\mathbf{c}_n$ 的長度（即 $\mathbf{c}_n$ 包含的字元數），將觀測狀態個數 M_n 設置爲合法漢字的個數。例如，合法熟語「天南地北」的 HMM 模型含有 $N_n = 4$ 個隱藏狀態，分別爲「天」、「南」、「地」、「北」；考慮漢字 Unicode 編碼區間 0x4E00 至 0x9FA5，共有 20902 個合法漢字，對應 $M_n = 20902$ 個觀測狀態。

其次，我們考慮三種熟語錯誤來源，即漢字重複、缺失和錯字。例如，將「天南地北」誤用爲「天南地地北」，是漢字重複錯誤；將「天南地北」誤用爲「天南北」，是漢字缺失錯誤；將「天南地北」誤用爲「田南地北」，是漢字錯字錯誤。[2] 進一步地，重複錯誤和缺失錯誤使用轉移機率矩陣 $\mathbf{A}_n$ 表示，錯字錯誤則使用混淆機率矩陣 $\mathbf{B}_n$ 表示。

具體而言，$[\mathbf{A}_n]_{i,i}$ 表示合法熟語 $\mathbf{c}_n$ 的第 i 個字元發生重複錯誤的機率，$[\mathbf{A}_n]_{i,i+1}$ 表示合法熟語 $\mathbf{c}_n$ 的第 i 個字元正確地轉移到第 $i+1$ 個字元的機率，而 $[\mathbf{A}_n]_{i,i+k}$ 表示合法熟語 $\mathbf{c}_n$ 的第 $i+1$ 至 $i+k-1$ 個字元（$k>1$）發生缺失錯誤的機率；$[\mathbf{B}_n]_{i,j}$ 則表示合法熟語 $\mathbf{c}_n$ 的第 i 個字元被使用爲第 j 個合法漢字的機率。當然，假設合法熟語 $\mathbf{c}_n$ 的第 i 個字元是合法漢字集合中的第 i' 個漢字，則 $[\mathbf{B}_n]_{i,i'}$ 是該字

[1] 可根據實際需要使用其它最優準則。例如，若改用 $\mathbf{c}_{\text{opt}} = \arg\max_n \Pr\{\lambda_n|\mathbf{w}\}$，可根據 Bayes Formula，將其等效地表示爲 $\mathbf{c}_{\text{opt}} = \arg\max_n \Pr\{\mathbf{w}|\lambda_n\}\Pr\{\mathbf{c}_n\}$，此時只需在式 (2) 的基礎上，加入先驗機率知識 $\Pr\{\mathbf{c}_n\}\,(n = 1, 2, \cdots, N_{\text{idiom}})$，即每個合法熟語在漢語環境下被使用到的機率（可用語料中出現的頻次表示）。本文以式 (2) 的最優準則爲例，進行闡述和展示。

[2] 值得說明的是，本文所提出的自動糾錯方法，對於重複、缺失和錯字等錯誤出現不止一次，甚至多種錯誤同時出現的複雜情況（如將「天南地北」誤用爲「天天南地地北」、「田楠帝北」、「天南帝北北北」等），都能夠予以糾正。詳見第4.1節「基本功能的檢驗和展示」。

元未出現錯誤的正確機率。矩陣 $\mathbf{A}_n$ 和 $\mathbf{B}_n$ 的行和（row-sum）都爲 1。

需要強調的是，爲了得到 $\mathbf{A}_n$、$\mathbf{B}_n$ 和 $\boldsymbol{\pi}_n$，理論上需要對熟語 $\mathbf{c}_n$ 的正確及錯誤使用情況進行大量統計。然而，大規模訓練資料的獲取難度極大。事實上，可以預期的是，即使僅對 $\mathbf{A}_n$、$\mathbf{B}_n$ 和 $\boldsymbol{\pi}_n$ 所有元素作合理的手工設置，本方法仍能表現出優秀的自動糾錯性能。我們將在第3.2節詳細說明各合法熟語 HMM 模型參數的賦值思路，並在第4章中結合實驗結果展開討論。本文所提出基於 HMM 模型的中文熟語自動糾錯實施方案如圖3所示。

3 實驗設置

3.1 資料集

本方法在構建自動糾錯模型時，需要爲每一個合法熟語建立各自的 HMM 模型，進而計算各 HMM 模型產生待糾錯熟語的後驗機率，最終根據機率值的排序給出糾正建議。因此，需要構建合法熟語表（作爲糾錯參考）、漢字混淆集（涵蓋近形字、同音字、近音字等，用於 HMM 模型參數設置）、熟語錯誤案例集（用於測試方法性能）等。具體如下：

1) 在構建合法熟語表時，我們使用 THUOCL 資料集 (Han et al., 2016) 中的成語詞庫作爲合法熟語表，如圖4所示。THUOCL 詞庫是由 Zhiyuan Liu 團隊整理推出的中文詞庫。該成語詞庫包含 8519 個成語，具有較高的代表性。

2) 在構建漢字混淆集時，已有 nlp-hanzi-similar 近形字[3]和 SimilarCharacter 同音字語料庫[4]等資料可供使用。此外，我們還自行構建了近音字語料庫。在構建近音字語料庫過程中，需用到「漢字轉拼音」和「拼音轉所有漢字」的基本操作，而該任務可以基於開放漢語字典-現代漢語字音資料庫[5]來實現。

3) 在測試本方法性能時，需要搜集具有代表性的熟語使用錯誤案例集。現有翰霖文教機構[6]、上海外國語大學[7]、揚州大學[8]、上海外國語大學附屬浦東外國語學校[9]等整理的資料，以及《多功能實用成語典》(Cai, 2016) 和《成語糾錯手冊》(Gao and Liu, 2011) 等著作可供使用。

[3]https://github.com/houbb/nlp-hanzi-similar
[4]https://github.com/contr4l/SimilarCharacter
[5]https://github.com/kfcd/hyzd
[6]https://www.han-lin.tw/chinese-form/
[7]http://www.newoaa.shisu.edu.cn/cc/cf/c6349a117967/page.htm
[8]https://jwc.yzu.edu.cn/info/1054/1902.htm
[9]https://www.msshw.pudong-edu.sh.cn/list/36/11580.html

Figure 4: THUOCL 中文詞庫，截圖自：http://thuocl.thunlp.org/

3.2 HMM 模型參數設置

如前文所述，理論上需要對各合法熟語 $\{\mathbf{c}_n \forall n\}$ 的使用情況進行大量統計，從而訓練出各合法熟語 HMM 模型參數 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n \forall n\}$。不過，此操作需要消耗大量資源，現實中很難實現。可以預期的是，即使僅對 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n \forall n\}$ 進行簡單合理的手工設置，本方法仍能表現出優秀性能。本節將介紹具體的設置思路和方法。

3.2.1 轉移機率矩陣 $\{\mathbf{A}_n \forall n\}$

如無其他先驗知識，在本文中，長度相等的合法熟語使用相同的轉移機率矩陣。例如，對於長度爲 $N = 6$ 的合法熟語「有志者事竟成」，其轉移機率矩陣 $\mathbf{A}$ 可簡單設置爲

$$\mathbf{A} = \begin{bmatrix} 0.10 & 0.80 & 0.07 & 0.03 & 0 & 0 \\ 0 & 0.10 & 0.80 & 0.07 & 0.03 & 0 \\ 0 & 0 & 0.10 & 0.80 & 0.07 & 0.03 \\ 0 & 0 & 0 & 0.10 & 0.80 & 0.10 \\ 0 & 0 & 0 & 0 & 0.10 & 0.90 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

在式 (3) 中，矩陣 $\mathbf{A}$ 由一個基本機率質量函數（probability mass function，PMF）進行移位（shifted）而產生。我們將該基本 PMF 記爲 $\mathbf{a}_0$。因此，式 (3) 中的 $\mathbf{a}_0$ 是一個有效長度爲 $L_{\mathbf{A}} = 4$ 的行向量（row vector），其取值爲

$$\mathbf{a}_0 = [0.10, 0.80, 0.07, 0.03]. \tag{4}$$

式 (4) 表示該合法熟語每個漢字的「正確轉移機率」都爲 0.80，漢字重複機率都爲 0.10，缺失 1 個漢字的機率爲 0.07，缺失 2 個漢字的機率爲 0.03，而缺失更多漢字的機率均爲

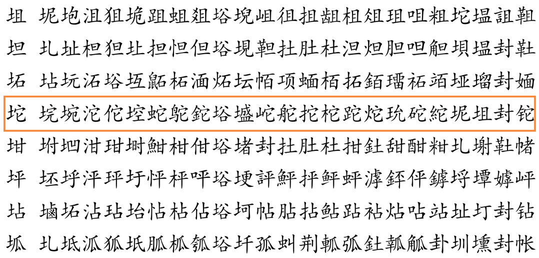


Figure 5: nlp-hanzi-similar 語料庫 (Hou, 2025) 的部分資料（該語料庫用於對近形字錯誤進行建模），整理自：https://github.com/houbb/nlp-hanzi-similar

0。[10]同時，矩陣 $\mathbf{A}$ 的下對角線元素（entries strictly below the main diagonal）均設爲 0，這是由於假設使用該合法熟語時發生漢字位置調換的機率小到可以忽略。當然，使用者可以根據實際情況，對矩陣 $\mathbf{A}$ 的基本 PMF 進行修改或學習。

3.2.2 觀測機率矩陣 $\{\mathbf{B}_n\,\forall n\}$

觀測機率矩陣 $\mathbf{B}_n$ 各元素的值表示各合法漢字錯用爲其它合法漢字的機率。同樣地，理論上需要利用漢字錯用資料進行統計學習，但這需要大量的計算資源，且相關資料並不充分，關於漢字錯用類型的分類也尚無統一的規範 (Liu et al., 2008)，現有資料難以直接利用。

鑒於此，本文提出一種簡單有效的觀測機率矩陣設置方法。首先，利用現有語料庫，爲每個合法漢字找到其對應的所有近形字、同音字和近音字。需要說明的是，「音」是指拼音，而在演算法實現時，可根據實際需求使用相應的拼音系統，例如漢語拼音、注音符號第二式、威妥瑪拼音、通用拼音等。本文以漢語拼音爲例進行闡述和實現。進而，將錯用機率在其近形字、同音字、近音字和所有其它合法漢字中進行分配。具體操作如下：

1) 由近形字導致的錯用情況，使用 nlp-hanzi-similar 語料庫 (Hou, 2025) 進行設置。例如，在圖5中，對於漢字「坨」，考慮「垞」、「埦」、「沱」、「佗」等近形字。

2) 由同音字導致的錯用情況，使用 SimilarCharacter 語料庫 (XiaoFang et al., 2025) 進行設置。例如，在圖6第 1599 行中，對於漢字「乖」，考慮「拐」、「怪」、「摑」共 3 個同音字[11]。

Figure 6: SimilarCharacter 語料庫 (XiaoFang et al., 2025) 的部分資料（該語料庫用於對同音字錯誤進行建模），截圖自：https://github.com/contr4l/SimilarCharacter

3) 對於由近音字導致的錯用情況，我們構建了近音字語料庫。在本文已完成的簡單原型系統中，暫時只考慮由前後鼻音以及平翹舌音導致的近音字錯誤。例如，拼音爲「can」的漢字，其近音字包括拼音爲「cang」和「chan」的所有漢字；拼音爲「an」的漢字，其近音字僅包括拼音爲「ang」的所有漢字。[12]對此，我們需要獲取每個漢字的拼音以及獲取各相近拼音對應的所有漢字，即實現「漢字轉拼音」和「拼音轉所有漢字」。

最後，結合上述近形字、同音字和近音字錯誤資料，可對各合法漢字錯用爲其它合法漢字的觀測機率值進行設置。令漢字使用的正確率爲 $p_{\mathbf{B}}$，則需要將錯用機率 $1-p_{\mathbf{B}}$ 在除該漢字之外的所有合法漢字中進行分配。假設近形字錯誤、同音字錯誤、近音字錯誤和其它錯誤等共計 4 種錯誤所佔比例分別爲 x_1、x_2、x_3 和 x_4，滿足 $x_1+x_2+x_3+x_4=100\%$。若無其它先驗知識，一種簡單有效的做法是：將近形字錯誤機率 $(1-p_{\mathbf{B}})x_1$、同音字錯誤機率 $(1-p_{\mathbf{B}})x_2$、近音字錯誤機率 $(1-p_{\mathbf{B}})x_3$ 和其它錯誤機率 $(1-p_{\mathbf{B}})x_4$ 分別在所對應的所有合法漢字中平均分配，最終得到觀測機率矩陣各元素的值。例如，當 $p_{\mathbf{B}}=0.90$，且 $x_1=20\%$、$x_2=40\%$、$x_3=30\%$ 和 $x_4=10\%$ 時，將近形字錯誤機率 $0.10x_1=0.02$、同音字錯誤機率 $0.10x_2=0.04$、近音字錯誤機率 $0.10x_3=0.03$ 和其它錯誤機率 $0.10x_4=0.01$ 分別在對應漢字中平均分配。

[10] 矩陣 $\mathbf{A}$ 的第 $N-L_{\mathbf{A}}+1$ 至 N 行的上對角線元素（entries strictly above the main diagonal）需要進行修正，以保證行和（row-sum）均爲 1，如式 (3) 中的第 3 至 6 行所示。

[11] 在本文中，爲方便起見，「同音字錯誤」是指拼音相同但音調可能不同的情形（如「廣」和「逛」也視爲同音字），並不局限於拼音和音調都相同。

[12] 後續工作中，使用者可以根據實際情況，對近音字錯誤類型進行修改和擴充，例如定義「nan」和「lan」爲相近拼音。

錯誤熟語	錯誤類型及個數	**Word** 檢查器	**GPT-5**	本方法
風中之濁	近形 1	風中之燭	風中之燭	風中之燭
逢場做戲	同音 1	逢場作戲	逢場作戲	逢場作戲
發號司令	近音 1	無糾錯建議	發號施令	發號施令
爭先後	缺字 1	無糾錯建議	爭先恐後	爭先恐後
更新換換代代	多餘 2	無糾錯建議	更新換代	更新換代
屋裡去鬧	同音 3	無糾錯建議	無理取鬧	無理取鬧
瘋馬牛兒不想記	同音 3、多餘 1	無糾錯建議	風馬牛不相及	風馬牛不相及
掛狗頭買羊肉	錯字 2、同音 1	無糾錯建議	掛羊頭賣狗肉	掛羊頭賣狗肉
畢奇工夫於一夜	同音 2、多餘 1、錯字 1	無糾錯建議	畢其功於一役	畢其功於一役
意點異低	同音 3	無糾錯建議	疑點重重	一點一滴
天婚棣岸	同音 3	無糾錯建議	天崩地裂	天昏地暗
有四五孔	同音 2、近音 1	無糾錯建議	有恃無恐	有恃無恐
溜繩吳組	同音 2、近音 2	無糾錯建議	無功受祿	六神無主
億民金仍	同音 2、近音 2	無糾錯建議	一鳴驚人	一鳴驚人
拔山涉稅	同音 2	無糾錯建議	拔山涉水	跋山涉水
图挠蜈蚣	同音 3、錯字 1	無糾錯建議	獨木難支	徒勞無功
一洋囊進	同音 1、近音 2	無糾錯建議	一網打盡	一言難盡
滸滸森威	同音 2、近音 1	無糾錯建議	虎虎生威	虎虎生威
喜夾良李	錯字 1、近音 1	無糾錯建議	喜出望外	喜結連理
恭喜花柴	錯字 1、近音 1	無糾錯建議	恭喜發財	恭喜發財

Table 1: 不同出錯嚴重程度下的中文熟語糾錯結果對比示例

測試集	樣本數	準確 糾正數	糾正 準確率
G&L	1285	1241	96.6%
YZU	104	101	97.1%
SISU	47	46	97.9%
SISU-PD	183	175	95.6%
總計	1619	1563	96.5%

Table 2: 使用測試集對本方法糾錯能力進行評估

3.2.3 初始機率向量 $\{\boldsymbol{\pi}_n \forall n\}$

在本文的建模思路中，$\boldsymbol{\pi}_n$ 的第 i 個元素表示人們在使用合法熟語 $\mathbf{c}_n$ 時，首個漢字是 $\mathbf{c}_n$ 的第 i 個漢字的機率。通常情況下，$\boldsymbol{\pi}_n$ 的第 1 個元素應相對較大，即正確機率相對較大。在後續實驗中，如無其他先驗知識，我們將 $\boldsymbol{\pi}_n$ 的第 1 個元素都設置爲 $[\boldsymbol{\pi}_n]_{1,1} = \pi_1$，其餘第 $j\,(j > 1)$ 個元素設置爲 $[\boldsymbol{\pi}_n]_{1,j} = \frac{1-\pi_1}{N_n-1}$。

3.3 評估準則

我們首先在第4.1節完成基本功能測試實驗，然後在第4.2和4.3節使用測試集進行全面評估。測試集包含多個測試樣本，每個測試樣本包含 1 個「錯誤寫法」和 1 個「正確寫法」。在測試實驗中，以每個測試樣本中的「錯誤寫法」作爲輸入，取本方法給出的機率最大的 N_{cand} 個糾錯建議作爲輸出，並與測試樣本中的「正確寫法」作比較。若至少 1 個糾錯建議與「正確寫法」相同，則視爲「準確糾正」，否則視爲「錯誤糾正」。統計各測試集的準確糾正數和錯誤糾正數，以準確糾正數與測試樣本數之比爲性能指標，稱糾正準確率（correction rate，CR）。

4 實驗結果及討論

4.1 基本功能的檢驗和展示

在本節中，我們以出錯程度不同的熟語作爲測試輸入，取本方法所給出的最佳糾錯建議作爲測試輸出（$N_{\text{cand}} = 1$），與微軟公司 Word 辦公軟體拼寫檢查器以及 GPT-5 的智能糾正結果進行比較，以檢驗本方法的基本功能。我們在不同測試案例中加入不同數量的近形字、同音字、近音字、缺失、多餘等錯誤類型，甚至有多種錯誤同時出現的情況。

在實驗中，所有轉移機率矩陣的基本 PMF 都設置爲 $\mathbf{a}_0 = [0.10, 0.80, 0.07, 0.03]$，漢字正確使用機率 $p_{\mathbf{B}} = 0.90$，$x_1 = 20\%$，$x_2 = 40\%$，$x_3 = 30\%$，$x_4 = 10\%$，初始機率向量 $\pi_1 = 0.9$，$\boldsymbol{\pi}_n = [\pi_1, \frac{1-\pi_1}{N_n-1}, \cdots, \frac{1-\pi_1}{N_n-1}]$。實驗結果如表1所示。[13]

[13]對於 GPT-5，所使用的提示詞爲：「風中之濁、逢場做戲、發號司令、爭先後、更新換換代代、屋裡去鬧、瘋馬牛兒不想記、掛狗頭買羊肉、畢奇工夫於一夜、意點異低、天婚棣岸、有四五孔、溜繩吳組、億民

序號	測試樣本中 的錯誤寫法	本方法 糾正建議	測試樣本中 的正確寫法	說明或啓示
1	固步自封	固步自封	故步自封	「固步自封」並非公認的正確用法。合法 熟語表的專業性、嚴謹性和無歧義性十分重要。
2	流言非語	流言飛語	流言蜚語	「流言飛語」和「流言蜚語」是異形詞。合法 熟語表的專業性、嚴謹性和無歧義性十分重要。
3	辛辛學子	過河卒子	莘莘學子	後續在完善漢字混淆集時， 需對易錯字情況進行更全面的考慮。
4	發人深醒	發人深思	發人深省	後續在完善漢字混淆集時， 需對一字多音的情況加以考慮。

Table 3: 表2中 SISU 和 YZU 兩個測試集的 4 個錯誤糾正測試樣本的細節及討論

由表1可見，Word 拼寫檢查器的糾錯能力最差。如前文所述，這是因爲該檢查器使用了自定義詞典，只能糾正已收錄於詞典中的特定錯誤（例如「風中之濁」和「逢場做戲」這兩個錯誤詞），而對於其它錯誤詞則無法處理。與之不同的是，本方法發揮了 HMM 模型在序列分析和機率建模方面的優勢，能夠靈活地糾正多樣化類型的熟語使用錯誤。

有趣的是，GPT-5 無法對一些錯誤熟語予以正確糾正，尤其是同音錯誤和近音錯誤比較嚴重時。例如，在表1中，GPT-5 建議將「意點異低」糾正爲「疑點重重」而非「一點一滴」，將「天婚棣岸」糾正爲「天崩地裂」而非「天昏地暗」，將「一洋囊進」糾正爲「一網打盡」而非「一言難盡」。同時，GPT-5 有時會給出錯誤的糾正建議，即糾正結果並不是公認的合法熟語，例如表1中的「拔山涉水」。相反，本方法模型簡單、參數少，無需複雜的訓練過程，仍能表現出與大語言模型相當（甚至更優）的熟語糾錯性能。當然，我們的目的不在於貶低 GPT-5 的熟語糾錯能力（畢竟熟語糾錯並非其唯一關注點），而是將其作爲對比基準之一，驗證本方法的有效性和靈活性。

4.2 使用測試集對系統性能進行評估

本節評估本方法的糾正准確率。採用鳳凰出版社《成語糾錯手冊》、揚州大學、上海外國語大學和上海外國語大學附屬浦東外國語學校等整理發佈的共計四個常見成語使用錯誤案例集作爲測試集，下文分別稱爲 G&L、YZU、SISU 和 SISU-PD 測試集。這四個測試集分別有 1285、104、47 和 183 個測試樣本，共計 1619 個。每個測試樣本包含 1 個「錯誤寫法」和 1 個「正確寫法」，例如錯誤的「直接了當」和正確的「直截了當」。同時，在測試實驗中，若測試樣本的「正確寫法」事先未收錄於 THUOCL 成語庫中，則將其加入並更新合法成語表，以完善合法成語表的收錄質量。

在實驗中，轉移機率矩陣基本 PMF（即 $\mathbf{a}_0$）、漢字正確使用機率 $p_{\mathbf{B}}$、漢字錯誤使用機率分配比例 (x_1, x_2, x_3, x_4)、初始機率向量 $\boldsymbol{\pi}_n$、糾正建議個數 N_{cand} 的設置都與第4.1節相同。實驗結果如表2所示。由表2可見，即使僅對 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n \forall n\}$ 進行簡單合理的手工設置，本方法對四個測試集的糾正準確率分別高達 96.6%、97.1%、97.9% 和 95.6%，總計 96.5%，因而在統計意義上具有優秀的糾錯能力。這說明本方法對熟語出錯過程的建模，以及對合法熟語 HMM 模型的轉移機率和觀測機率等參數的設置及相應物理意義的理解，都是合理且有效的。

同時，爲了更好地理解本方法的特點，我們將表2中 SISU 和 YZU 兩個測試集的 4 個錯誤糾正測試樣本記錄下來，如表3所示，並作討論如下：

1) 第 1 個樣本中，本方法認爲「固步自封」一詞無錯誤。該測試樣本之所以未準確糾正，是由於實驗所使用的 THUOCL 成語表收錄了「固步自封」一詞。因此，該錯誤糾正與本方法的核心技術無關。當然，這也給本文後續工作帶來啓示：合法熟語表的專業性、嚴謹性和無歧義性十分重要。

2) 第 2 個樣本中，本方法將「流言非語」糾正爲「流言飛語」，而測試樣本的正確寫法是「流言蜚語」。經查閱資料，「流言飛語」和「流言蜚語」是一組異形詞 (Wang et al., 2001)。然而，THUOCL 成語表收錄了「流言飛語」一詞。因此，與上一個樣本相同，該錯誤糾正與本方法的核心技術無關。

3) 第 3 個樣本中，本方法未將「辛辛學子」糾正爲「莘莘學子」。這是因爲在構建漢字混淆集時，未充分考慮「莘」字容易被錯用爲「辛」字這一情況，即未將「莘」和「辛」視爲

金仍、拔山涉稅、図挠蜈蚣、一洋囊進、滸滸森威、一萬无寄、恭喜花柴，以上這些詞，分別最可能是哪個熟語的錯誤使用？」網址爲：https://chatgpt.com/。

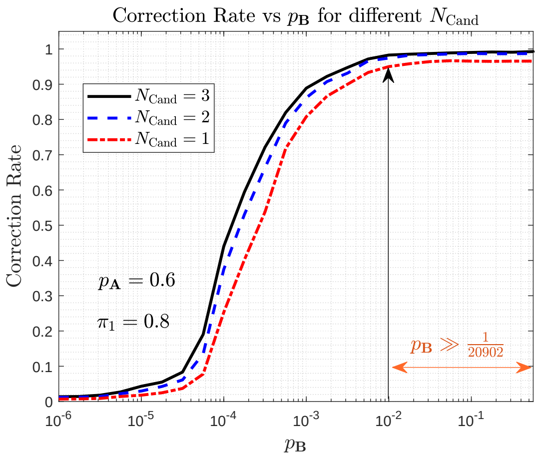


Figure 7: 正確糾正率受觀測機率矩陣取值影響的實驗結果圖，其中 N_{cand} 表示本方法輸出的糾正建議個數

近形字或近音字。儘管這個錯誤糾正與本方法的核心建模思路無關，但也啓示我們：後續需對易錯字情況進行更全面的考慮，進一步完善漢字混淆集的構建。

4) 第 4 個樣本中，本方法未將「發人深醒」糾正爲「發人深省」。這是因爲在構建同音字混淆集時，只考慮「省」字的讀音「sheng」，未考慮「省」字的另一個讀音「xing」，即未將「省」和「醒」視爲同音字。儘管這個錯誤糾正與本方法的核心建模思路無關，但也啓示我們：後續需對一字多音的情況加以考慮，進一步完善漢字混淆集的構建。

4.3 健壯性測試

在本文中，我們並未嘗試通過收集每一個合法熟語 $\{\mathbf{c}_n \forall n\}$ 的正確及錯誤使用情況（因其代價巨大）來訓練每一個 HMM 模型的參數 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n \forall n\}$。有趣的是，由第4.1和4.2節可見，本方法只需對 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n \forall n\}$ 進行簡單合理的手工設置，就能得到良好的糾正準確率。事實上，由於使用了機率模型而非編輯距離模型，並且各合法熟語之間天然具有較明顯的差別，本方法對於 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n \forall n\}$ 的不同取值具有很強的健壯性（robustness）。對此，我們完成以下實驗進行驗證。

實驗採用第4.2節中的四個測試集共計 1619 個測試樣本進行。同時，對於長度爲 N_n 的合法熟語 $\mathbf{c}_n$，其對應的 HMM 模型參數設置如下：狀態轉移矩陣 $\mathbf{A}_n$ 的基本 PMF 爲 $\mathbf{a}_0 = [\frac{1-p_A}{3}, p_A, \frac{1-p_A}{3}, \frac{1-p_A}{3}]$；混淆機率矩陣 $\mathbf{B}_n$ 中，$x_1 = x_2 = x_3 = 30\%$，$x_4 = 10\%$；初始機率向量設置爲 $\boldsymbol{\pi}_n = [\pi_1, \frac{1-\pi_1}{N_n-1}, \cdots, \frac{1-\pi_1}{N_n-1}]$。

首先，我們考察觀測機率矩陣 $\{\mathbf{B}_n \forall n\}$ 中的漢字正確使用機率 p_B 對糾正準確率的影響，實驗結果如圖7所示。由圖可見，只要 p_B 取值明顯大於 $\frac{1}{20902}$，本方法都能給出高於 95% 的糾正准確率，呈現出關於觀測機率矩陣的健壯性。[14] 另外，糾正准確率隨著輸出建議個數 N_{cand} 的增加而提升。有趣的是，即便只取機率最大的唯一一個糾錯建議作爲輸出（$N_{\text{cand}} = 1$），在 $p_B > 0.01$ 時也能獲得高於 95% 的糾正准確率。最後，我們考察轉移機率矩陣基本 PMF 中的 p_A 和初始機率向量的 π_1 值對糾正准確率的影響，實驗結果如表4和5所示。可以看出，本方法對於 $\{\mathbf{A}_n \forall n\}$ 和 $\{\boldsymbol{\pi}_n \forall n\}$ 的不同取值同樣具有很強的健壯性。以上特點有助於降低收集大量熟語使用情況進行模型參數學習的必要性。

p_A	0.4	0.6	0.8	1
$\text{CR}_{0.01}$	94.6%	94.9%	95.3%	94.3%
$\text{CR}_{0.1}$	96.7%	96.7%	96.9%	95.4%

Table 4: 糾正準確率受轉移機率矩陣取值影響的實驗結果，其中 CR_x 表示 $p_B = x$ 時的結果

π_1	0.4	0.6	0.8	1
$\text{CR}_{0.01}$	95.1%	95.3%	95.2%	95.1%
$\text{CR}_{0.1}$	96.9%	96.9%	96.8%	96.6%

Table 5: 糾正準確率受初始機率向量取值影響的實驗結果，其中 CR_x 表示 $p_B = x$ 時的結果

5 結論

本文提出了一種輕量型的中文熟語自動糾錯新方法。通過構建包含近形字、同音字和近音字的漢字混淆集，以及利用現有的合法術語庫，將中文熟語糾錯問題建模爲隱藏式馬可夫模型的基本問題並求解。相比於傳統的編輯距離和自定義詞典方法，本方法模型參數少、計算簡單，即使僅對模型參數進行簡單合理的人工設置，也能獲得出色的糾錯性能，對出錯程度嚴重的熟語予以準確糾正。同時，本方法對模型參數具有很強的健壯性，因而無需耗費大量資源用於訓練資料獲取和模型參數學習。在後續的工作中，需要完善合法熟語表和漢字混淆集的構建，以提升模型的糾正準確性。此外，可進一步研究如何推廣到考慮上下文資訊的場景，以增大本方法的適用范圍。

[14] $p_B > \frac{1}{20902}$ 在實際使用中很容易滿足。此處的 20902 是觀測狀態個數，即合法漢字個數。此外，p_B 不能嚴格等於 1，因爲 $p_B = 1$ 會導致 $\Pr\{\mathbf{w}|\lambda_n\} = 0 \forall n$，從而導致異常結果。

References

Zongyang Cai. 2016. 多功能實用成語典（第二版）*(Multi-functional Practical Idioms Dictionary (Second Edition)) [in Chinese]*. Taipei: Wu-Nan Book Inc.

Yulin Gao and Peishu Liu. 2011. 成語糾錯手冊 *(Idiom Correction Manual) [in Chinese]*. Nanjing: Phoenix Publishing and Media Inc.

Shiyi Han, Yuhui Zhang, Yunshan Ma, Cunchao Tu, Zhipeng Guo, Zhiyuan Liu, and Maosun Sun. 2016. THUOCL：Tsinghua Open Chinese Lexicon. Website. `http://thuocl.thunlp.org/`.

Victoria J. Hodge and Jim Austin. 2003. A comparison of standard spell checking algorithms and a novel binary neural approach. *IEEE Transactions on Knowledge and Data Engineering*, 15(5):1073–1081.

Binbin Hou. 2025. 漢字相似度計算工具及中文形近字算法 (Chinese character similarity calculation tools and algorithms for Chinese similar-looking characters) [in Chinese]. Website. `https://github.com/houbb/nlp-hanzi-similar`.

Amita Jain and Minni Jain. 2014. Detection and correction of non word spelling errors in Hindi language. In *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*, pages 1–5.

Dan Jurafsky and James H. Martin. 2024. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (Third edition draft). Website. `https://web.stanford.edu/~jurafsky/slp3/`.

Muhammad Ifte Khairul Islam, Rahnuma Islam Meem, Faisal Bin Abul Kasem, Aniruddha Rakshit, and Md. Tarek Habib. 2019. Bangla spell checking and correction using edit distance. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–4.

Weihui Lin, Weijun Tang, Xiaofan Lin, Rongbin Zhang, Shaojie Xu, Xiaojuan Ning, and Wenting Zhang. 2012. *Hidden Markov model and its application in typewriting correction.* Project report in the course Machine Learning and Its Applications, South China University of Technology, Guangzhou.

Chao-Lin Liu, Chih-Bin Huang, Juei-Yu Weng, and Yi-Hsuan Chuang. 2008. 形音相近的易混淆漢字的搜尋與應用 (Identification and applications of visually confusing Chinese characters) [in Chinese]. In *Proceedings of the 20th Conference on Computational Linguistics and Speech Processing (ROCLING)*, pages 108–122.

Microsoft-Corporation. 2025. Add or edit words in a spell check dictionary. Website. `https://support.microsoft.com/`.

Ei Phyu Phyu Mon, Ye Kyaw Thu, Than Than Yu, and Aye Wai Oo. 2021. SymSpell4Burmese: Symmetric delete spelling correction algorithm (SymSpell) for Burmese spelling checking. In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6.

Peter Norvig. 2016. How to write a spelling corrector. Website. `http://norvig.com/spell-correct.html`.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

A. Revathi, M. Vimaladevi, and N. Arivazhagan. 2023. Spelling correction using encoder-decoder and Damerau-Levenshtein distance. In *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, pages 469–472.

Moch Yusup Soleh and Ayu Purwarianti. 2011. A non word error spell checker for Indonesian using morphologically analyzer and HMM. In *Proceedings of the International Conference on Electrical Engineering and Informatics*, pages 1–6.

Mark Stamp. 2021. A revealing introduction to hidden Markov models. Website. `https://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf`.

Kseniia Varlamova, Ildar Khabutdinov, and Andrey Grabovoy. 2023. Automatic spelling correction for Russian: Multiple error approach. In *2023 Ivannikov Ispras Open Conference (ISPRAS)*, pages 169–175.

Jihong Wang, Ming Chen, and Liqing Ren. 2001. 現代實用漢語字典 *(Modern Practical Chinese Dictionary) [in Chinese]*. Shanghai: Shanghai Far East Publishers.

Ziqi Wang, Gu Xu, Hang Li, and Ming Zhang. 2014. A probabilistic approach to string transformation. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1063–1075.

XiaoFang, mystical001, and demo. 2025. 對常用的 6700 個漢字進行音、形比較，輸出音近字、形近字的列表 (Compare the sounds and shapes of 6,700 commonly used Chinese characters and output a list of characters with similar sounds and shapes) [in Chinese]. Website. `https://github.com/contr4l/SimilarCharacter`.

Chengqing Zong. 2024. Lecture Notes on Natural Language Understanding, Chinese Academy of Sciences, Beijing. Website. `https://nlpr.ia.ac.cn/cip/ZongReportandLecture/ReportandLectureIndex.htm`.

邁向繁體中文 ModernBERT 的初步研究
Toward Traditional Chinese ModernBERT: A Preliminary Study

Yi-En Chen, Qiao-Ying He, Kuan-Yu Chen
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
yechen.research@gmail.com, qiaoyinghe.cs@gmail.com, kychen@mail.ntust.edu.tw

摘要

本研究使用多項最新技術，包含 RoPE、Flash Attention 等，並在結合大規模中文網路語料與百科全書資料的基礎上，預訓練一個專爲長文本設計的繁體中文編碼器模型。我們在閱讀理解、文本分類等任務上進行評估，結果顯示模型效能整體落後於既有中文基準。透過 pseudo-perplexity 分析，我們推測預訓練階段未能充分學習資料分布，並討論了超參數、收斂與資料品質等可能影響因素。雖然結果不理想，本研究仍提供中文語言模型發展的實驗經驗與改進方向。

Abstract

This study employs several state-of-the-art techniques, including RoPE and Flash Attention, and leverages large-scale Chinese web corpora and encyclopedic data to pre-train an encoder model specifically designed for long text in Traditional Chinese. We evaluate the model on tasks such as reading comprehension and text classification, and the results show that its overall performance lags behind existing Chinese benchmarks. Through pseudo-perplexity analysis, we infer that the pre-training phase did not sufficiently capture the data distribution, potentially due to factors such as hyperparameters, convergence, and data quality. Although the results are suboptimal, this study still offers valuable experimental insights and directions for improving Chinese language model development.

關鍵字： Transformer、繁體中文、長文本、預訓練語言模型

Keywords: Transformer, Traditional Chinese, Long Context, Pretrained Language Model

1 前言 Introduction

自 BERT (Devlin et al., 2019) 問世以來，僅編碼器（encoder-only）的 Transformer (Vaswani et al., 2017) 模型已成爲眾多自然語言處理（NLP）任務的核心架構。儘管近年來主流的大型語言模型（LLM）如 GPT (Radford et al., 2018, 2019)、Qwen (Bai et al., 2023; Yang et al., 2024) 等多採用僅解碼器（decoder-only）架構，但基於編碼器的模型在性能與計算效率的平衡上仍具備獨特優勢，並在文本分類、命名實體識別（NER）、資訊檢索（IR）等下游任務中獲得廣泛應用。

上下文長度對大型語言模型而言至關重要，擁有更長的上下文處理能力，使 LLM 能夠勝任更多樣化的任務。同樣地，對於 BERT、RoBERTa 等編碼器模型，上下文長度亦是關鍵因素。然而，這些基於 Transformer 的模型由於 self-attention 機制具有 $O(n^2)$ 的時間複雜度，限制了其處理長上下文的能力，導致可處理的上下文長度普遍較短。

雖然後續推出的編碼器模型在上下文長度方面有所改進，但主要集中於英文和簡體中文的優化，針對繁體中文的研究相對稀少。因此，本研究旨在填補此一研究缺口，爲繁體中文的長上下文編碼器模型發展貢獻力量。

2 研究方法 Methods

2.1 模型架構 Model Architecture

本研究使用了多項已經過廣泛測試的最新進展，以提升模型表現以及訓練效率:

- Rotary Positional Embedding (RoPE)：相較於傳統的絕對位置編碼，RoPE (Su et al., 2024) 能更有效捕捉長距離依賴，提升模型處理長文本的能力。
- Pre-Normalization： 採用 Pre-Normalization (Xiong et al., 2020) 結構，有助於穩定訓練過程，促進深層模型的收斂。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 133–139
November 20-22, 2025 ©2025 Association for Computational Linguistics

- Alternating Global and Local Attention：交替使用全局與局部注意力機制 (Beltagy et al., 2020)。在局部注意力下，token 僅能與 sliding window 內的其他 token 互動；全局注意力則可與整個序列互動。我們每三層採用一次全局注意力。局部注意力的 RoPE theta 設爲 10,000，全局注意力的 RoPE theta 設爲 160,000。

- Unpadding：將 batch 內所有序列的 padding token 移除 (Zeng et al., 2022) 並串接成一個長序列計算，減少不必要的運算，提升訓練效率。

- Flash Attention：使用 Flash Attention (Dao et al., 2022) 以降低記憶體使用量，提升訓練速度。

2.2 中文分詞器 Tokenizer

本研究使用 benchang1110/Qwen2.5-Taiwan-1.5B-Instruct [1] 所提供的分詞器做爲基礎，此分詞器基於 Qwen2.5 (Qwen Team, 2024) 模型，並透過 tokenizer swapping，將簡體中文的 token 替換爲相對應的繁體中文 token，以優化對繁體中文的處理能力。

且爲了符合 BERT 模型的相容性，我們在此額外加入了 `[PAD]`、`[UNK]`、`[SEP]`、`[CLS]`、`[MASK]` 等特殊 token，並將模型配置中的 pad_token 明確指定爲 `[PAD]`，以取代原有的 <|endoftext|>，使模型的可以沿用以前 BERT 的預訓練任務。

2.3 訓練資料 Training Data

本研究的預訓練資料主要由以下兩部分構成:

- **FineWeb2**: FineWeb2 [2] 是一個大規模、開放且多語言的資料集，專門爲訓練高品質的大型語言模型 (LLM) 而設計。其資料主要源自於對 Common Crawl 存檔的大規模處理與精煉。該資料集涵蓋了從 2013 年到 2024 年 4 月的 96 個 Common Crawl 數據快照，並透過一個包含過濾、去重和語言辨識的複雜流程進行處理，旨在爲開發能夠理解和生成多種語言文字的大型語言模型提供一個強大且可擴展的高品質資源。我們使用的是 FineWeb2 的中文部分 (cmn_Hani) 的一個子集，約包含 4 千萬個樣本，其中繁體中文的樣本比例較高。

- 中文維基百科資料: 爲了增強模型在中文領域的知識與理解能力，我們額外納入了完整的中文維基百科資料 [3]。此資料集包含了百科全書中的所有條目，內容涵蓋歷史、文化、科技、地理等多個領域，具有結構化、高品質、事實性強的特點。使用此資料有助於模型學習中文世界豐富的背景知識和正規的書面語表達方式。其中約包含 140 萬個樣本。

2.4 資料預處理 Data Preprocessing

爲了確保訓練資料的品質與針對性，我們對原始資料進行了以下預處理步驟：

- 內容過濾：我們移除了訓練資料中的非必要章節，包括但不限於參考資料、外部連結、延伸閱讀等章節。這些章節通常包含大量的 URL 連結、引用格式等非自然語言內容，對模型學習語言理解能力的幫助有限，反而可能引入雜訊。

- 地區與語言變體篩選：考慮到中文存在不同的地區變體（如台灣、中國大陸、香港等），且不同地區在用詞、表達習慣上存在差異，我們在資料篩選時以保留臺灣正體的部分爲主，並賦予其較高的優先度。

經過上述預處理後，最終用於預訓練的資料集共包含約 35.7B 個 tokens，在保持多樣性的同時，也具備了更高的針對性與品質。

2.5 預訓練任務 Pretraining Task

本研究採用遮罩語言模型 **(Masked Language Modeling, MLM)** 任務，並結合 n-gram masking 策略，在中文語料上進行模型預訓練。整體訓練過程分爲以下兩個的階段：

1. 第一階段 (短文本學習): 模型以最大序列長度 1024 的文本進行訓練。此階段的核心目標是讓模型掌握中文的基礎語法結構與核心語義知識。

2. 第二階段 (長文本建模): 在繼承第一階段學習到的權重基礎上，我們將最大序列長度擴展至 8192 進行接續訓練。此階段旨在顯著提升模型對長距離文本依賴關係的捕捉與建模能力。

詳細的超參數設置如表 1所示，模型訓練使用 8 張 NVIDIA H100 GPU，總訓練時間約爲 7 天。

[1] https://huggingface.co/benchang1110/Qwen2.5-Taiwan-1.5B-Instruct

[2] https://huggingface.co/datasets/HuggingFaceFW/fineweb-2

[3] https://dumps.wikimedia.org/zhwiki/20250520/

	Phase 1	Phase 2
Max Sequence Length	1024	8192
Batch Size	512	512
Training Steps	233000	40000
Learning Rate	8e-4	3e-4
LR Schedule	WSD	WSD
Weight Decay	1e-5	1e-5
Optimizer	AdamW	AdamW
Betas	(0.9, 0.999)	(0.9, 0.999)
Epsilon	1e-8	1e-8

表 1. 訓練參數設定

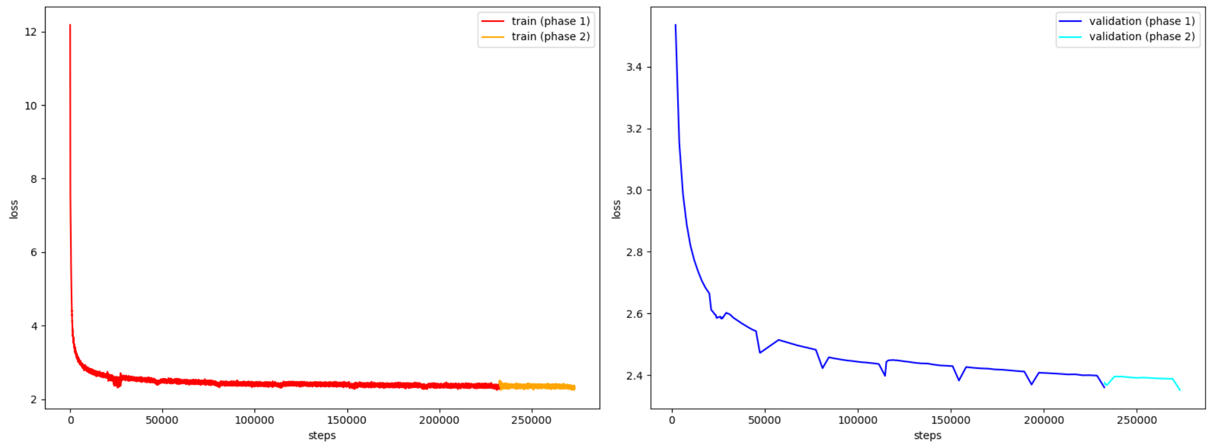


圖 1: 預訓練過程中的損失函數變化。左圖爲訓練損失，右圖爲驗證損失。

3 評估 evaluation

我們將模型在多個下游任務上進行實驗，以驗證其有效性，並將結果與其他研究論文中報告的中文模型結果進行比較 (Cui et al., 2021) 。實驗任務包括：

- 閱讀理解任務：CMRC2018 (Cui et al., 2019)、DRCD (Shao et al., 2019)
- 單句分類任務：ChnSentiCorp (Tan and Zhang, 2008)、THUCNews (Li and Sun, 2007)、TNEWS (Xu et al., 2020)
- 句子對分類任務：XNLI (Conneau et al., 2018)、LCQMC (Liu et al., 2018)、BQ (Chen et al., 2018)、OCNLI (Hu et al., 2020)

由於訓練語料中繁體中文比例較高，所有資料集在實驗前皆使用 OpenCC [4] 進行簡繁轉換。

[4]https://github.com/BYVoid/OpenCC

3.1 閱讀理解任務 Machine Reading Comprehension

閱讀理解任務會提供模型一組 (文本, 問題) 的文字對，模型需根據文本內容生成問題的答案，答案通常爲文本中的一段文字（span）。我們使用 DRCD 與 CMRC2018 兩個中文閱讀理解資料集進行評估。結果展示在表 2. 中，其中 EM (Exact Match) 表示完全匹配率，F1 Score 表示 F1 分數。從結果可見，本研究模型在 DRCD 資料集上的表現與早期 BERT 模型（如 BERT-base）相當，但在 CMRC2018 資料集上則有顯著的效能落差，尤其在測試集（Test set）的表現遠不如其他基準模型，這可能暗示模型在特定領域的泛化能力較弱。

3.2 單句分類任務 Single Sentence Classification

單句分類任務會提供模型一段句子，要求模型根據該句子進行分類。我們在三個中文資料集上進行驗證：

- ChnSentiCorp：中文情感分析資料集，包含正向（positive）與負向（negative）兩個類別。

	CMRC2018				DRCD			
	Dev		Test		Dev		Test	
Model	EM	F1	EM	F1	EM	F1	EM	F1
BERT-base	65.5	84.5	70.0	87.0	83.1	89.9	82.2	89.2
BERT-wwm	66.3	85.6	70.5	87.4	84.3	90.5	82.8	89.7
BERT-wwm-ext	67.1	85.7	71.4	87.7	85.0	91.2	83.6	90.4
RoBERTa-wwm-ext	67.4	87.2	72.6	89.4	86.6	92.5	85.6	92.0
ELECTRA-base	68.4	84.8	73.1	87.1	87.5	92.5	86.9	91.8
MacBERT-base	68.5	87.9	73.2	89.5	89.4	94.3	89.5	93.8
Ours	55.2	81.3	32.2	68.1	83.4	92.5	82.6	91.9

表 2. 各模型於 CMRC2018 及 DRCD 資料集之閱讀理解任務表現，(單位：%)。

- THUCNews：新聞分類資料集，本研究使用其子集，任務爲判斷輸入新聞文本的主題類別。
- TNEWS：中文短新聞分類資料集，包含 15 個類別，任務爲對輸入文本進行主題分類。

表 3. 展示了各模型在單句分類任務上的準確率 (Accuracy) 表現。在 ChnSentiCorp 與 TNEWS 任務上，本模型表現與基準模型相近；然而，在 THUCNews 新聞分類任務上，效能則有明顯落後，顯示模型在處理長文本或特定主題分類時可能存在不足。

3.3 句子對分類任務 Sentence Pair Classification

句子對分類任務會提供模型一組 (文本, 文本) 的文字對，要求模型根據該句子進行分類。我們在四個中文資料集上進行驗證：

- XNLI: 跨語言自然語言推理資料集，本研究使用中文部分，任務爲判斷句子對是否具備「蘊涵、矛盾或中立」關係（textual entailment）。
- LCQMC: 大規模中文問答語料庫，用於評估中文問答對語意相似度的自然語言處理資料集。
- BQ Corpus: 中文語義等價判斷（Sentence Semantic Equivalence Identification, SSEI）資料集，用於識別句子對是否語義相同。
- OCNLI: 原生中文自然語言推理資料集，爲首個非翻譯、專爲中文語境設計的大型 NLI 資料集。

表 4 展示了各模型在句子對分類任務上的準確率 (Accuracy) 表現。相較於其他任務，本模型在句子對分類的整體表現較不理想。雖然在 XNLI 與 LCQMC 任務上僅略遜於基準模型，但在 BQ Corpus 與 OCNLI 資料集上出現了更大幅度的效能下降，尤其是在 OCNLI 這個專爲中文原生語境設計的資料集上，表現最不理想，這可能反映出模型在捕捉語意相似度與自然語言推論的細微差異方面仍有待加強。

4 討論 Discussion

雖然模型在多個下游任務微調後，表現仍顯著落後於現有基準，我們推測主要原因來自預訓練階段未能充分學習語料的潛在分佈。

爲了量化模型對訓練語料的擬合程度，我們引入了僞困惑度（Pseudo-Perplexity, PPPL）(Salazar et al., 2020) 作爲評估指標。對於自回歸語言模型，較低的困惑度（PPL）通常意味著模型能更好地建模語料；而 PPPL 則可對遮罩語言模型（MLM）進行近似評估。對於一個語料集 $\mathcal{W}$，其 PPPL 的計算方式如下：

$$\mathrm{PPPL}(\mathcal{W}) := \exp\left(-\frac{1}{N}\sum_{\mathbf{W}\in\mathcal{W}} \mathrm{PLL}(\mathbf{W})\right)$$

其中 N 是語料集的總詞數，而 $\mathrm{PLL}(\mathbf{W})$ 是單一樣本 $\mathbf{W}$ 的僞對數似然率（Pseudo-Log-Likelihood），其定義爲：

$$\mathrm{PLL}(\mathbf{W}) := \sum_{t=1}^{|\mathbf{W}|} \log P_{\mathrm{MLM}}(W_t \mid \mathbf{W}_{\setminus t}; \Theta) \quad (1)$$

此處模型需預測每個詞 w_t 在其上下文 $\mathbf{w}_{\setminus t}$ 中的機率。

我們從訓練語料中隨機採樣文本，並在本模型及幾個基準模型上計算 pseudo-perplexity。結果如表 5. 顯示，部分基準模型對這些文本表現出更低的 pseudo-perplexity 值，這表明本模型在預訓練階段可能未能充分捕捉資料分布。基於此，我們提出以下幾個可能原因：

Model	ChnSentiCorp Dev	ChnSentiCorp Test	THUCNews Dev	THUCNews Test	TNEWS Dev
BERT-base	94.7	95.0	97.7	97.8	56.3
BERT-wwm	95.1	95.4	98.0	97.8	56.5
BERT-wwm-ext	95.4	95.3	97.7	97.7	57.0
RoBERTa-wwm-ext	95.0	95.6	98.3	97.8	57.4
ELECTRA-base	93.8	94.5	98.1	97.8	56.1
MacBERT-base	95.2	95.6	98.2	97.7	57.4
Ours	94.4	95.0	93.0	93.8	56.2

表 3. 各模型於 ChnSentiCorp、THUCNews 及 TNEWS 資料集之單句分類任務表現，(單位：%)。

Model	XNLI Dev	XNLI Test	LCQMC Dev	LCQMC Test	BQ Corpus Dev	BQ Corpus Test	OCNLI Dev
BERT-base	77.8	77.8	89.4	86.9	86.0	84.8	86.0
BERT-wwm	79.0	78.2	89.4	87.0	86.1	85.2	86.1
BERT-wwm-ext	79.4	78.7	89.6	87.1	86.4	85.3	86.4
RoBERTa-wwm-ext	80.0	78.8	89.0	86.4	86.0	85.0	86.0
ELECTRA-base	77.9	78.4	90.2	87.6	84.8	84.5	84.8
MacBERT-base	80.3	79.3	89.5	87.0	86.0	85.2	86.0
Ours	77.1	77.1	86.0	86.6	82.0	79.6	70.0

表 4. 各模型於 XNLI、LCQMC、BQ Corpus 及 OCNLI 資料集之句子對分類任務表現，(單位：%)。

Model	pppl
BERT-base	2.49
BERT-wwm	2.73
BERT-wwm-ext	3.48
MacBERT-base	13.39
Ours	5.60

表 5. 各模型於本模型訓練集之 pseudo-perplexity。

4.1 超參數設置不理想

雖然大部分超參數與原始論文保持一致，但由於訓練語料不同，原論文的超參數對中文語料可能不完全適用，可能導致模型未達最佳效果。

4.2 模型尚未完全收斂

若模型在建模能力上仍不理想，可能原因之一是模型尚未完全收斂。我們推測第一階段短文本訓練可能過早結束，模型尚未充分學習資料分布。

4.3 訓練資料品質

資料中可能存在雜訊或標註不均衡，這會影響模型學習到準確的語言分布，進而影響下游任務表現。

除了預訓練因素，下游任務的資料分布亦可能影響模型表現。如同 2.3節所述，訓練語料以繁體中文爲主，因此模型的語言分布更接近繁體中文。然而下游任務資料集多爲簡體中文，即便經過繁簡轉換，語法、詞彙偏好與用字習慣仍存在差異，進而導致本模型在這些任務上的表現不如其他基準模型。

5 結論 Conclusion

在這篇論文中，我們嘗試使用文語料，對現有的 state-of-the-art 語言模型架構進行預訓練，並在多個下游任務上進行評估。雖然最終的實驗結果尚未能超越既有的中文基準模型，但這些結果爲我們提供了寶貴的觀察。

我們的分析指出，模型效能不佳可能與語料分布、語言特性、預訓練設定以及詞彙表設計等因素有關。這些推論顯示，將現有方法直接應用於中文語料並非一件簡單的工作，仍需更多針對中文語言特性的調整與研究。

在未來的工作中，我們計畫針對語料來源與品質進行優化，並嘗試更合適的預訓練策略與模型設定。我們相信這些經驗將成爲我們後續研究的重要基礎，幫助我們在中文語言模型的探索上持續改進。

6 致謝 Acknowledgment

This work was supported by the National Science and Technology Council (NSTC) of Taiwan under Grants NSTC 112-2636-E-011-002, NSTC 112-2628-E-011-008-MY3, and NSTC 113-2640-B-002-005. Additional support was provided by the ”Empower Vocational Education Research Center” at the National Taiwan University of Science and Technology (NTUST) through the Featured Areas Research Center Program, as part of the Higher Education Sprout Project funded by the Ministry of Education (MOE), Taiwan. The authors also thank the National Center for High-Performance Computing, National Applied Research Laboratories (NARLabs), Taiwan, for providing essential computational and storage resources.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and et al. 2023. Qwen technical report. Technical report, arXiv preprint arXiv:2309.16609.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. 2018. The BQ corpus: A large-scale domain-specific Chinese corpus for sentence semantic equivalence identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4946–4951, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504—3514.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: fast and memory-efficient exact attention with io-awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Lawrence S. Moss. 2020. Ocnli: Original chinese natural language inference.

Jingyang Li and Maosong Sun. 2007. Scalable term selection for text categorization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 774–782, Prague, Czech Republic. Association for Computational Linguistics.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. LCQMC:a large-scale Chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1952–1962, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2019. Drcd: a chinese machine reading comprehension dataset.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Songbo Tan and Jin Zhang. 2008. An empirical study of sentiment analysis for chinese documents. *Expert Syst. Appl.*, 34(4):2622—2629.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and et al. 2024. Qwen2 technical report. Technical report, arXiv preprint arXiv:2407.10671.

Jinle Zeng, Min Li, Zhihua Wu, Jiaqi Liu, Yuang Liu, Dianhai Yu, and Yanjun Ma. 2022. Boosting distributed training performance of the unpadded bert model.

Effective Speaker Diarization Leveraging Multi-task Logarithmic Loss Objectives

Jhih-Rong Guo[1], Tien-Hong Lo[1], Yu-Sheng Tsao[2], Pei-Ying Lee[1], Yung-Chang Hsu[2], and Berlin Chen[1]
[1]National Taiwan Normal University, Taiwan
[2]EZ-AI, Taiwan
{jhihrong, teinhonglo, 60947089s, berlin}@ntnu.edu.tw,
{sam, mic}@ez-ai.com.tw

Abstract

End-to-End Neural Diarization (EEND) has undergone substantial development, particularly with powerset classification methods that enhance performance but can exacerbate speaker confusion. To address this, we propose a novel training strategy that complements the standard cross entropy loss with an auxiliary ordinal log loss, guided by a distance matrix of speaker combinations. Our experiments reveal that while this approach yields significant relative improvements of 15.8% in false alarm rate and 10.0% in confusion error rate, it also uncovers a critical trade-off with an increased missed error rate. The primary contribution of this work is the identification and analysis of this trade-off, which stems from the model adopting a more conservative prediction strategy. This insight is crucial for designing more balanced and effective loss functions in speaker diarization.

Keywords: speaker diarization, powerset classification, loss function, ordinal log loss, Pyannote

1 Introduction

Speaker diarization is the task of determining ”who spoke when” in a recording with multi-speaker. Clustering-based (Wang et al., 2018; Landini et al., 2022; Garcia-Romero et al., 2017) are typically structured as a pipeline of modules, including Voice Activity Detection (VAD), speaker embedding extraction, and a clustering algorithm. While clustering-based approaches can evolve with advancements in speaker embedding and clustering algorithms, its inherent limitation of assigning only a single speaker to each frame still prevents it from performing well on overlapped speech.

Although some studies (Bullock et al., 2020; Charlet et al., 2013) have attempted to mitigate the inherent limitation of clustering-based by using methods such as Overlapped Speech Detection (OSD). However, the additional modules may exacerbate the problem of error propagation within the pipeline. To address the problem of overlapping speech, End-to-End Neural Diarization (EEND) (Fujita et al., 2019a,b; Horiguchi et al., 2020) was proposed. This approach trains a single neural network to directly output the diarization result, thereby removing the potential for error propagation. Furthermore, EEND formulates diarization as a multi-label classification task, which enables it to process overlapped speech. Nevertheless, its direct application to longer audio recordings is impractical due to memory requirements and degraded performance when handling more than four speakers.

The EEND-VC framework, introduced by Kinoshita et al. (2021), ingeniously integrates clustering-based with EEND, bypassing the challenges of standard EEND by applying the EEND model to shorter chunks. Nevertheless, a significant hurdle for most EEND-related methods is the immense amount of training data they require, typically requiring thousands of hours, which necessitates a dependency on simulated data. Consequently, the mismatch between these simulated datasets and the target domain typically requires further model adaptation. To enable training directly on real data, the Pyannote framework (Bredin, 2023) applies EEND to even shorter chunks, enabling the assumption that only a few speakers are present within each chunk. This approach significantly reduces the data dependency, making it feasible to train the EEND model directly on real data.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 140–145
November 20-22, 2025

Recent advancements building upon the Pyannote framework have delivered superior performance in speaker diarization. These improvements are largely attributed to key strategies such as switching speaker diarization from multi-label to powerset multi-class classification problem (Plaquet and Bredin, 2023) and leveraging pre-trained Self-Supervised Learning (SSL) models with more robust encoder like the Conformer (Han et al., 2025; Plaquet et al., 2025). However, while the powerset formulation offers significant advantages over multi-label methods, it can also exacerbate issues related to speaker confusion. Consequently, enhancing the ability of model to classify speakers accurately within powerset remains a valuable area for future research.

In this paper, we use the cross entropy loss as the main objective function and introduce an ordinal log loss (Castagnos et al., 2022) that considers distances between different classes as an auxiliary objective. Because we believe that using cross entropy loss alone makes it difficult for the model to learn the relationship between different classes during training (e.g., {1} and {1, 2} both contain speaker 1). Although speaker diarization is typically evaluated using nominal metrics, we contend that strategic incorporation of a distance-aware objective function can be beneficial. We call this hybrid objective function as the Multi-task Logarithmic Loss (Multi-task Log Loss). This combination has been proven effective in ordinal classification (Kasa et al., 2024).

2 Methodology

2.1 Multi-task Log Loss

Since speaker diarization is a task primarily evaluated using nominal metrics, we employ the cross entropy loss ($\mathcal{L}_{CE}$) as main objective function:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} p_i log(\hat{p}_i), \tag{1}$$

where N represents the number of classes, and p_i is 1 if class $i \in \{1, 2, ..., N\}$ is the ground-truth class and 0 otherwise. Assuming that class j is the ground-truth label, cross entropy loss can be simplified to $-log(\hat{p}_j)$, where $\hat{p}_j$ denotes the probability for class j as predicted by the model.

To guide the model to learn the relationships between different classes, we incorporate an ordinal log loss ($\mathcal{L}_{OLL}$) as an auxiliary objective function. This approach utilizes a distance matrix to define the distance between classes, where each class represents a unique combination of speakers. The loss is formulated as follows:

$$\mathcal{L}_{OLL} = -\sum_{i=1}^{N} log(1 - \hat{p}_i) d(j, i)^{\alpha}, \tag{2}$$

where $d(j, i)$ is the distance between class j and class i, which is defined by the distance matrix D and scaled by a hyperparameter α.

The multi-task log loss ($\mathcal{L}_{MLL}$) is composed of cross entropy loss and ordinal log loss:

$$\mathcal{L}_{MLL} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{OLL}, \tag{3}$$

where λ is a hyperparameter that respectively determine the weight of the contribution of ordinal log loss to the overall loss.

2.2 Distance Matrix

In ordinal tasks, a distance matrix can be readily constructed from explicitly defined relationships between classes. However, such ordinal relationships are absent in speaker diarization. Therefore, we propose to construct a distance matrix based on the set-theoretic relationships between the different speaker combinations. The distance D_{ij} between any two speaker sets, s_i and s_j, is defined by their symmetric difference, which counts the number of speakers present in one set but not the other. This can be formulated as the size of their union minus the size of their intersection:

$$D_{ij} = |s_i \cup s_j| - |s_i \cap s_j|, \tag{4}$$

where $S = \{s_1, s_2, ..., s_N\}$ represents the set of powerset classes. Assuming each segment contains $C = 3$ speakers and a maximum of $K = 2$ overlapping speakers, the number of powerset class is $N = 7$:

- $\emptyset$ for non-speech frames;
- $\{1\}$, $\{2\}$ and $\{3\}$ for one speaker;
- $\{1, 2\}$, $\{1, 3\}$ and $\{2, 3\}$ for two speaker.

For example, the distance between the class representing speakers 1 and 2, $s_i = \{1,2\}$, and the class representing only speaker 1, $s_j = \{1\}$, would be $D_{ij} = |\{1,2\} \cup \{1\}| - |\{1,2\} \cap \{1\}| = |\{1,2\}| - |\{1\}| = 2 - 1 = 1$. This intuitively means there is one speaker difference between the two classes. Therefore, when $C = 3$ and $K = 2$, the distance matrix D is:

$$D = \begin{pmatrix} 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 1 & 0 & 2 & 2 & 1 & 1 & 3 \\ 1 & 2 & 0 & 2 & 1 & 3 & 1 \\ 1 & 2 & 2 & 0 & 3 & 1 & 1 \\ 2 & 1 & 1 & 3 & 0 & 2 & 2 \\ 2 & 1 & 3 & 1 & 2 & 0 & 2 \\ 2 & 3 & 1 & 1 & 2 & 2 & 0 \end{pmatrix} \quad (5)$$

2.3 Speaker Diarization Pipeline

We adopt the same three-stage pipeline as Pyannote, which proceeds sequentially through three main components:

1. Segmentation: The input audio is first split into overlapping short segments, and End-to-End Neural Diarization (EEND) is applied to each segment to produce local diarization results.

2. Embedding: Based on the local diarization information, speaker embeddings are extracted from speech segments corresponding to each speaker.

3. Clustering: The extracted speaker embeddings are grouped using a clustering algorithm to map speakers across all segments and generate the final global speaker diarization result.

For the segmentation stage within our pipeline, we retrain the model by adopting the EEND framework proposed by Han et al. (2025). As depicted in Figure 1, the architecture first extracts features from an audio input using a pre-trained WavLM model. The feature outputs from each layer are subsequently combined through a weighted sum with learnable parameters to create a fused representation. This representation then undergo a projection layer and layer normalization before being fed into the Conformer. Finally, another linear layer as the classifier, producing logits for the N output classes. During training, all parameters of the pre-trained WavLM backbone are kept frozen.

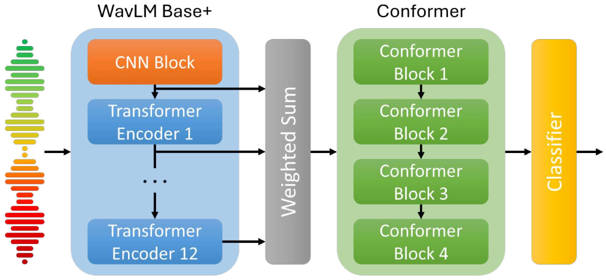


Figure 1: The architecture of EEND model.

3 Experiments

To ensure that experimental results are reproducible, we will conduct model training and evaluation using the DiariZen toolkit[1], which is driven by AudioZen and Pyannote 3.1.

3.1 Baseline

In this paper, we conduct a comparative analysis against a model trained exclusively with cross entropy loss. To ensure a fair evaluation, we maintained a consistent model architecture and configuration for all experiments, with the exception being the additional hyperparameters introduced by the multi-task log loss.

3.2 Datasets

We use AMI, AliMeeting, and AISHELL-4 as datasets for model training and evaluation, with total durations of 98.38, 126.34, and 120.25 hours respectively.The detailed duration for each dataset is presented in Table 1.

Table 1: A summary of the datasets (hrs.)

Dataset	Train	Dev	Test
AMI	79.65	9.67	9.06
AliMeeting	111.36	4.21	10.78
AISHELL-4	97.39	10.14	12.73
Compound	288.40	24.01	32.56

3.3 Evaluation Metrics

For evaluation, we employ Diarization Error Rate (DER), which is the sum of three error types: Missed Error Rate (MER), the percentage of speech time that is incorrectly labeled as non-speech; False Alarm Rate (FAR), the percentage of non-speech time incorrectly labeled as speech; and Confusion Error Rate (CER), the percentage of speech time assigned to the wrong speaker.

[1]https://github.com/BUTSpeechFIT/DiariZen

Table 2: A comparison of speaker diarization performance on the AMI, AliMeeting, and AISHELL-4 datasets for EEND model trained with cross entropy loss ($\mathcal{L}_{CE}$) versus multi-task log loss ($\mathcal{L}_{MLL}$).

	AMI			**AliMeeting**			**AISHELL-4**		
	MER	**FAR**	**CER**	**MER**	**FAR**	**CER**	**MER**	**FAR**	**CER**
$\mathcal{L}_{CE}$ **(baseline)**	**9.08**	3.94	4.46	**8.58**	3.07	7.13	**2.96**	4.29	3.41
- 250ms collar	6.87	1.95	2.58	4.54	0.87	5.63	1.21	1.71	2.36
$\mathcal{L}_{MLL}$ **(proposed)**	10.51	**3.12**	**4.03**	9.25	**2.55**	**6.30**	3.62	**3.79**	**3.16**
- 250ms collar	8.07	1.46	2.30	5.09	0.70	4.95	1.52	1.48	2.19

Table 3: An ablation study on the performance of the multi-task log loss with varying weights (λ). This comparison highlights three scenarios: (1) $\lambda = 0.5$, which yields the best performance, alongside (2) $\lambda = 0.3$ and (3) $\lambda = 0.7$, which represent cases with a lesser and greater influence from the ordinal log loss, respectively.

	Compound			
	DER	**MER**	**FAR**	**CER**
baseline	15.47	**6.65**	3.74	5.08
(1)	**15.23**	7.51	**3.15**	**4.57**
(2)	15.59	7.57	**3.15**	4.88
(3)	15.54	7.74	3.22	4.58

Table 4: A comparison of the performance with different distance between the non-speech and speaker-active classes within the distance matrix. The conditions are as follows: (1) represents the original configuration, (4) sets the distance to 2 (i.e., $d(\emptyset, i) = 2$ and $d(j, \emptyset) = 2$), and (5) sets the distance to 4.

	Compound			
	DER	**MER**	**FAR**	**CER**
baseline	15.47	**6.65**	3.74	5.08
(1)	**15.23**	7.51	**3.15**	**4.57**
(4)	15.69	6.95	3.72	5.02
(5)	15.80	6.80	3.82	5.19

3.4 Experimental Setups

The EEND model was trained on the compound training set and validated on the compound development set, using a pre-trained WavLM Base+ model[2] as a frozen feature extractor. We set the maximum number of speakers to $C = 4$ and the maximum overlapping speakers to $K = 2$. The input audio was divided into 8-second segments with a 6-second hop size. The model was trained for a maximum of 100 epochs using the AdamW optimizer with a learning rate of 1×10^{-3} and a batch size of 64. Early stopping with a patience of 10 epochs was applied based on the validation loss. The hyperparameter α for ordinal log loss was set to 1.5. In the subsequent diarization pipeline, speaker embeddings were extracted using the ResNet34-LM[3], followed by Agglomerative Hierarchical Clustering (AHC) to produce the final output.

[2]https://huggingface.co/microsoft/wavlm-base-plus

[3]https://huggingface.co/pyannote/wespeaker-voxceleb-resnet34-LM

3.5 Results

The diarization performance of the EEND models, trained with either the conventional cross entropy loss or our proposed multi-task log loss, is detailed in this section. Table 2 presents a comprehensive comparison of the two models across the AMI, AliMeeting, and AISHELL-4, evaluated under two conditions: with no forgiveness collar (rows 1 and rows 3) and with a 250ms forgiveness collar (rows 2 and rows 4). On the compound dataset, our proposed multi-task log loss achieves relative improvements of 15.8% in FAR and 10.0% in CER compared to the baseline. These gains, however, are accompanied by a notable regression in MER, which leads to only a marginal improvement in the overall DER from 15.47% to 15.23%. This outcome suggests a fundamental trade-off: our ordinal-aware loss function effectively guides the model to be more precise in identifying speakers and avoiding false speech detection, but it does so by adopting a more conservative behavior. We will further explain this in subsequent experiments.

To determine the optimal contribution of our auxiliary objective, we conducted an ablation study on its weight, λ, with results shown in Table 3. The findings indicate that a weight of $\lambda = 0.5$ yields the best overall DER. While different weights modulate the balance between MER, FAR, and CER, the study reinforces the previously observed trade-off, where a lower FAR and CER consistently correlate with a higher MER compared to the baseline.

We hypothesize that in segments of high uncertainty, the model prefers to predict non-speech to minimize the penalties associated with incorrect speaker assignments, thus increasing the MER. To further investigate the cause of the elevated MER and validate our hypothesis regarding the model's conservative behavior, we performed a targeted analysis by modifying the distance between the non-speech class and all speaker-active classes in the distance matrix. The results, presented in Table 4, reveal a clear and direct relationship. As the distance from the non-speech class is increased (conditions (4) and (5)), the MER shows a corresponding improvement. However, this improvement comes at the cost of a gradual regression in both FAR and CER. This experiment confirms our hypothesis: a smaller distance incentivizes the model to predict non-speech in uncertain segments as a low-penalty alternative, leading to more missed errors. Conversely, a larger distance forces the model to make more definitive—and consequently, more error-prone—classifications among speaker-active classes.

In summary, our investigation into applying an ordinal-aware loss to the EEND framework has yielded a crucial insight. While the proposed multi-task log loss effectively reduces FAR and CER, its primary contribution is the revelation of a distinct trade-off with the MER. Our experiments, particularly the analysis of the non-speech class distance, provide strong evidence that this trade-off arises directly from the incentive of model to adopt a more conservative prediction strategy under this loss structure. Therefore, the key takeaway from our results is the characterization of this complex behavior. This insight is critical for understanding the implications of incorporating ordinal constraints in powerset speaker diarization.

4 Conclusion

In this paper, we investigated the effect of introducing an ordinal log loss to the training of an EEND model. Our findings demonstrate that equipping the model with distance information between different speaker combination classes effectively enhances performance in terms of FAR and CER, yielding relative improvements of 15.8% and 10.0%, respectively. However, these gains were largely offset by a regression in the MER, which resulted in only a marginal improvement in the overall DER. We further identified that this MER degradation was directly linked to the distance assigned to the non-speech class within our proposed distance matrix. Our experiments confirmed that a smaller distance incentivizes the model to adopt a more conservative prediction strategy in uncertain segments, thereby increasing missed speech errors. Therefore, the key takeaway from our results is the identification and explanation of this complex behavior. This insight is critical for understanding the implications of incorporating ordinal constraints in powerset-based speaker diarization and offers a clear direction for future improvements.

5 Future Work

Based on our findings, we propose two potential improvements. First, the manually defined, set-theoretic distance matrix could be replaced by a data-driven approach. A future direction would be to learn the distances between speaker combination classes directly from the training data. This could yield a distance matrix that is more optimally aligned with the acoustic features of the data and potentially improve the overall balance of the proposed multi-task log loss. Second, to directly counteract the MER regression observed in our experiments, we propose integrating feature fusion techniques that have proven effective for VAD. Inspired by recent findings from Tripathi et al. (2025), who demonstrated that fusing traditional MFCC features with pre-trained model representations can significantly reduce the MER, we plan to explore a similar strategy. A promising approach would be to incorporate a feature fusion module at the input stage of our EEND model.

References

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. Interspeech*.

Latané Bullock, Hervé Bredin, and Leibny Paola Garcia-Perera. 2020. Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection. In *Proc. ICASSP*.

François Castagnos, Martin Mihelich, and Charles Dognin. 2022. A simple log-based loss function for ordinal text classification. In *Proc. COLING*.

Delphine Charlet, Claude Barras, and Jean-Sylvain Liénard. 2013. Impact of overlapping speech detection on speaker diarization for broadcast news and debates. In *Proc. ICASSP*.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. 2019a. End-to-end neural speaker diarization with permutation-free objectives. In *Proc. Interspeech*.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019b. End-to-end neural speaker diarization with self-attention. In *Proc. ASRU*.

Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree. 2017. Speaker diarization using deep neural network embeddings. In *Proc. ICASSP*.

Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukáš Burget. 2025. Leveraging self-supervised learning for speaker diarization. In *Proc. ICASSP*.

Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu. 2020. End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors. In *Proc. Interspeech*.

Aniket Kasa, Siva Rajesh Goel, Karan Gupta, Sumegh Roychowdhury, Pattisapu Priyatam, Anish Bhanushali, and Prasanna Srinivasa Murthy. 2024. Exploring ordinality in text classification: A comparative study of explicit and implicit techniques. In *Proc. ACL*.

Keisuke Kinoshita, Marc Delcroix, and Naohiro Tawara. 2021. Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds. In *Proc. ICASSP*.

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: Theory, implementation and analysis on standard tasks. *Computer Speech & Language*.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. Interspeech*.

Alexis Plaquet, Naohiro Tawara, Marc Delcroix, Shota Horiguchi, Atsushi Ando, Shoko Araki, and Hervé Bredin. 2025. Dissecting the segmentation model of end-to-end diarization with vector clustering.

Kumud Tripathi, Chowdam Venkata Kumar, and Pankaj Wasnik. 2025. Attention is not always the answer: Optimizing voice activity detection with simple feature fusion. In *Proc. Interspeech*.

Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. 2018. Speaker diarization with lstm. In *Proc. ICASSP*.

Leveraging Weak Segment Labels for Robust Automated Speaking Assessment in Read-Aloud Tasks

Yue-Yang He, Berlin Chen
National Taiwan Normal University, Taipei, Taiwan
{yueyanghe, berlin}@ntnu.edu.tw

Abstract

Automated speaking assessment (ASA) has become a crucial component in computer-assisted language learning, providing scalable, objective, and timely feedback to second-language learners. While early ASA systems relied on hand-crafted features and shallow classifiers, recent advances in self-supervised learning (SSL) have enabled richer representations for both text and speech, improving assessment accuracy. Despite these advances, challenges remain in evaluating long speech responses, due to limited labeled data, class imbalance, and the importance of pronunciation clarity and fluency, especially for read-aloud tasks. In this work, we propose a segment-based ASA framework leveraging WhisperX to split long responses into shorter fragments, generate weak labels from holistic scores, and aggregate segment-level predictions to obtain final proficiency scores. Experiments on the GEPT corpus demonstrate that our framework outperforms baseline holistic models, generalizes robustly to unseen prompts and speakers, and provides diagnostic insights at both segment and response levels.

Keywords: Automated Speaking Assessment, WhisperX, Weak Labels

1 Introduction

With the rapid advances in computing technology and the growing population of second-language (L2) learners, automated speaking assessment (ASA) has attracted increasing attention and become an essential component in computer-assisted language learning (CALL). ASA systems are designed to provide timely and reliable feedback on learners' speaking performance, enabling them to improve their oral proficiency in an autonomous and low-stress environment. In addition, ASA offers scalable, objective, and consistent evaluations, thereby alleviating the workload of language instructors and facilitating large-scale language learning applications.

Early ASA research primarily relied on shallow classifiers and hand-crafted features that captured different aspects of speaking competence, such as delivery (e.g., pronunciation, fluency, intonation), content (e.g., appropriateness, relevance), and language use (e.g., grammar, vocabulary) (Cucchiarini et al., 1998; Chen et al., 2010; Coutinho et al., 2016; Chen et al., 2018; Qian et al., 2019; Wu et al., 2022). More recently, the emergence of self-supervised learning (SSL) paradigms has opened up new opportunities for ASA. Text-based SSL models, such as BERT and its derivatives (Devlin et al., 2019), provide contextualized embeddings that have been successfully adopted in various language assessment tasks, including sentence-level evaluation (Arase et al., 2022), essay scoring (Nadeem et al., 2019; Wu et al., 2023), and spoken monologue assessment (Craighead et al., 2020). In parallel, the rapid development of speech-based SSL models, such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2023), has further strengthened ASA systems by offering rich acoustic representations (Bannò and Matassoni, 2023; McKnight et al., 2023; Wu and Chen, 2024; Lo et al., 2024).

Despite these advances, automated speaking assessment still faces persistent challenges in handling long speech responses. A representative example is the read-aloud task, where learners are evaluated primarily on pronunciation clarity and fluency. While text-based

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 146–152

models can capture lexical accuracy, they are inherently limited in assessing these speech-specific aspects. Moreover, the development of reliable ASA systems is hindered by the scarcity of large-scale annotated data, as existing datasets are often limited in size and imbalanced across proficiency levels. The computational cost of processing extended speech recordings further compounds these difficulties. Consequently, the lack of sufficient labeled resources restricts model robustness and limits the ability to deliver fine-grained and diagnostic feedback.

In this work, we explore an ASA framework designed to address both the scarcity of labeled data and the challenges of long speech recordings. Specifically, we leverage WhisperX (Bain et al., 2023) to process long audio responses and obtain time-aligned segments, each of which is subsequently evaluated with segment-level scoring. To compensate for the lack of labeled resources, we weakly associate each segment with the holistic proficiency score of the full response, thereby generating weak labels for training. This strategy not only increases the number of training instances, especially for underrepresented proficiency levels, but also highlights weaker segments where learner performance diverges from holistic expectations. Finally, segment-level predictions are aggregated (e.g., by mean or median) to reconstruct the overall proficiency score, offering a straightforward and interpretable mapping from local to global assessment.

Experiments on the GEPT corpus demonstrate that our framework consistently outperforms baseline holistic models and generalizes robustly to unseen prompts and speakers. We also investigate whether partial scoring of only the first or last 30 seconds of speech can approximate holistic judgments, revealing systematic differences that highlight both strengths and limitations of segment-level scoring.

In summary, our contributions are threefold:

1. We introduce a segment-based ASA framework for long read-aloud tasks that alleviates the scarcity of sentence-level annotations by exploiting weak labels derived from holistic scores;
2. We examine aggregation strategies for mapping segment-level predictions to holistic scores; and
3. We provide a comprehensive analysis of ASR quality and response-length effects on ASA performance.

These results offer new insights for designing ASA systems that are both data-efficient and diagnostically informative.

2 Related Work

2.1 Evolution of Automated Speaking Assessment Systems

Research on automated speaking assessment (ASA) has evolved from traditional feature engineering to the adoption of deep neural architectures. Early approaches relied on shallow classifiers with hand-crafted features targeting specific dimensions of proficiency, such as pronunciation, fluency, prosody, grammar, and vocabulary (Cucchiarini et al., 1998; Chen et al., 2010; Coutinho et al., 2016). While such systems demonstrated the feasibility of automatic scoring, their performance was often constrained by the limited representational power of manually designed features.

The advent of self-supervised learning (SSL) has substantially advanced ASA. On the text side, models such as BERT (Devlin et al., 2019) provide contextualized embeddings that have been successfully applied to various assessment tasks, including essay scoring (Nadeem et al., 2019), readability estimation (Arase et al., 2022), and spoken monologue evaluation (Craighead et al., 2020). These approaches leverage the semantic and syntactic richness of pre-trained language models, enabling more robust prediction of learner proficiency. In parallel, speech-based SSL models, such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2023), have emerged as powerful tools for capturing acoustic and phonetic information. Recent studies demonstrate their effectiveness in proficiency prediction and related tasks (Bannò and Matassoni, 2023; McKnight et al., 2023; Lo et al., 2024), showing that such representations can encode both linguistic and paralinguistic aspects critical to ASA.

However, most existing ASA systems treat each spoken response as a single, monolithic input, which becomes increasingly problematic when applied to long read-aloud tasks. Long-form speech raises both computational and temporal costs during training and inference, and more importantly, such systems typically produce only a holistic score without revealing which specific portions of the response contributed to the learner's performance. As a result, localized feedback is largely absent, and the literature contains relatively little work explicitly targeting the unique challenges of long-form ASA.

2.2 Handling Long Audio Inputs by WhisperX

WhisperX (Bain et al., 2023) is a system designed to efficiently transcribe long-form audio with word-level timestamps. It utilizes Voice Activity Detection (VAD) to segment audio into approximately 30-second chunks, which are then transcribed in parallel by Whisper and aligned with phoneme recognition models to produce accurate word-level timestamps. This approach enables batched inference, resulting in a twelve-fold speedup without sacrificing transcription quality. The segmentation process reduces issues like hallucinations and repetition, and the forced alignment ensures time-accurate transcriptions, making WhisperX suitable for applications such as subtitling and diarization.

3 Methodology

In this section, we describe the overall pipeline of our proposed Automated Speaking Assessment (ASA) framework, as illustrated in Figure 1. The system processes long audio responses by dividing them into manageable fragments, scoring each fragment independently, and subsequently aggregating these scores into a single holistic proficiency score.

3.1 Segmentation

Each spoken response in our dataset lasts approximately 90 seconds, which poses challenges for both ASR accuracy and downstream scoring. To address this, we employ WhisperX to obtain word-level timestamps. These timestamps allow us to segment each recording into shorter, coherent units of speech, hereafter referred to as *segments*. Each segment contains a contiguous portion of the learner's response, providing a finer-grained basis for subsequent scoring.

3.2 Weak-label Assumption

Since human raters typically provide only one holistic score per response, no ground-truth labels exist at the segment level. To overcome this limitation, we adopt a weak supervision strategy by assigning the holistic score of the full response to each of its segments as a weak-label. While this assumption may introduce label noise—because individual segments may not fully reflect overall proficiency—it substantially increases the number of training instances and enables finer-grained analysis of learner performance. This trade-off is particularly valuable under our limited-data setting.

3.3 Segment-Level Scoring

Each audio segment is processed independently to enable segment-level assessment. The Whisper encoder is adopted as the acoustic backbone, and its representations are fed into a grader module trained with weak segment-level supervision derived from holistic scores. This architecture effectively enlarges the usable training distribution, especially for low-resource proficiency levels, while providing localized diagnostic feedback that would otherwise be lost under holistic-only scoring.

3.4 Aggregation Strategies

Finally, the system aggregates segment-level predictions into a holistic proficiency score for the entire response. We consider multiple strategies, including simple averaging and median pooling, to examine which approach best captures the relationship between localized performance and the overall judgment. Moreover, variations among segment scores can highlight weaker portions of a response, offering diagnostic information beyond the final holistic score.

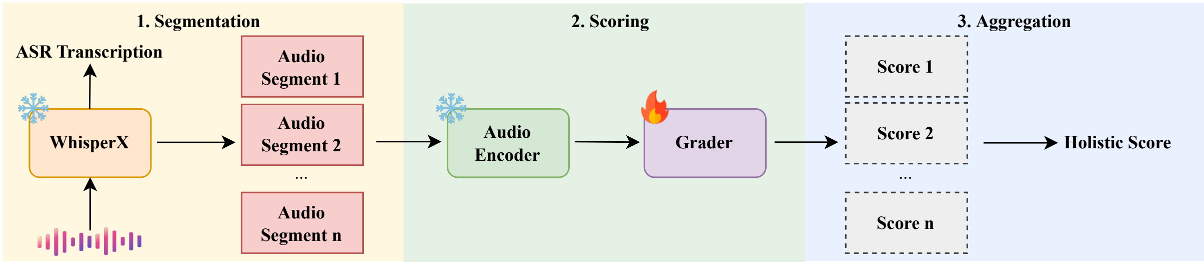


Figure 1: Proposed ASA framework: long read-aloud responses are segmented, each segment is scored independently, and the results are aggregated into a holistic proficiency score.

	1	2	3	4	5
Train	0	52	505	787	96
Valid	0	9	61	97	13
Known Content	0	6	67	99	8
Unknown Content	0	1	157	392	40

Table 1: Number of speakers for each holistic score in the GEPT dataset.

4 Experiments and Results

4.1 Dataset

This study utilizes a private corpus collected from the reading aloud task[1] in the General English Proficiency Test (GEPT), an important large-scale English assessment in Taiwan. In this task, participants were instructed to read aloud two given paragraphs within two minutes. The corpus consists of responses to eight different paragraph sets, with each set corresponding to a distinct passage.

Each response was independently scored by two professional raters on a five-point scale, where 1 represents the lowest performance and 5 the highest. The final score was obtained by averaging the two ratings. To evaluate model generalization, we define responses from unseen paragraph sets as the unknown content test set, while responses from previously seen sets are regarded as known content. The remaining data was further split into training, development, and test subsets following an 80/10/10 ratio.

The overall score distribution across training, validation, and test partitions is summarized in Table 1. This partitioning strategy ensures that the dataset supports evaluation under both familiar and novel content conditions, which is critical for assessing model robustness in practical applications.

4.2 Experimental Setup

We employed Whisper-large-v2[2] as our acoustic encoder in our framework. Model configurations were initialized using pretrained models from the HuggingFace Transformers library (Wolf et al., 2020). Training was conducted on a single NVIDIA 3090 GPU using Adam optimizer with a weight decay of 1e-5. The learning rate was set to 2e-4, and training was conducted for 15 epochs with a batch size of 25.

Baseline As baselines, we employed both a text-based SSL model and a speech-based SSL model, namely BERT[3] and wav2vec 2.0[4]. For the text-based baseline, the read-aloud audio was first transcribed by Whisper-large-v2, and the resulting text embeddings were extracted using a frozen BERT encoder; the same grading module used in our proposed framework was fine-tuned on top of it to predict holistic proficiency scores. For the speech-based baseline, we adopted wav2vec 2.0 as a frozen acoustic encoder and fine-tuned only the grading module on top of its representations.

Evaluation Metrics We evaluated model performance using three metrics: accuracy (ACC), weighted F1 score (F1), and Pearson correlation coefficient (PCC). ACC is defined as the proportion of predictions that exactly match the human-assigned holistic score. The

[1] https://www.gept.org.tw/Exam_Intro/t02_introduction.asp

[2] https://huggingface.co/openai/whisper-large-v2

[3] https://huggingface.co/google-bert/bert-base-uncased

[4] https://huggingface.co/facebook/wav2vec2-base

	Strategies	Known Content			Unknown Content		
		ACC↑	F1↑	PCC↑	ACC↑	F1↑	PCC↑
BERT	-	61.67	52.20	0.462	70.50	61.98	0.295
W2V	-	58.33	52.28	0.217	68.01	60.63	0.217
Whisper	First only	73.89	70.64	0.577	75.93	72.57	0.496
	Last only	78.33	74.97	0.679	76.10	72.72	0.499
Proposed	Mean	74.44	71.57	0.722	76.77	74.54	**0.623**
	Median	**82.22**	**79.04**	**0.748**	**78.47**	**76.01**	0.562

Table 2: Experimental results on the GEPT test dataset. "Known Content" denotes test samples with seen content, while "Unknown Content" denotes test samples with unseen content.

weighted F1 score accounts for label imbalance across proficiency levels, providing a more reliable estimate of performance on underrepresented categories. PCC further measures the monotonic relationship between predicted and reference scores, reflecting how well the model preserves the human-assigned ranking structure. These metrics jointly capture both discrete correctness (ACC) and ordinal consistency (F1, PCC), and are consistent with common practice in automated speaking assessment.

4.3 Results and Discussion

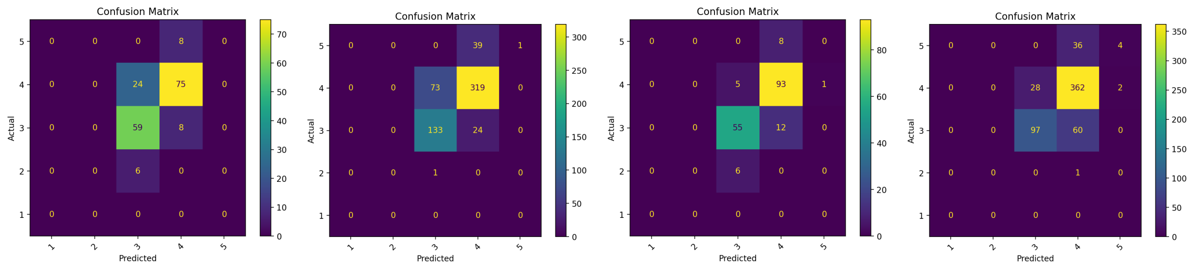
Baseline Performance Analysis. Table 2 summarizes the performance of our models under different configurations. The text-based baseline (BERT with Whisper transcription) achieved acceptable accuracy, highlighting the limitation of relying solely on ASR transcripts for holistic scoring. Interestingly, the speech-based SSL model (wav2vec 2.0) produced performance comparable to BERT in accuracy and weighted F1, but its PCC was substantially lower, particularly on the known-content set. This indicates that although both baselines can correctly classify a similar proportion of samples at the categorical level, the wav2vec-based model struggles to preserve the ordinal relationship among proficiency levels, likely due to its predictions being more distributionally concentrated and less sensitive to fine-grained prosodic variation relevant for human scoring. In contrast, BERT implicitly benefits from lexical cues captured via ASR, which may preserve a closer monotonic alignment with human-assigned proficiency levels.

Effect of Full-Length Training. The Whisper-based grader trained on full-length read-aloud recordings substantially outperformed both baselines across all three metrics, confirming the effectiveness of leveraging acoustic-prosodic information beyond lexical content. The performance gain in PCC further suggests that holistic fluency and speech quality are better reflected in continuous acoustic patterns than in discrete lexical sequences extracted from ASR transcriptions.

Temporal Coverage Analysis. To investigate the effect of temporal coverage, we compared models using only the first 30 seconds and the last 30 seconds of each recording. Both truncated variants yielded a noticeable drop across all metrics relative to the full-length model, suggesting that proficiency-related cues are distributed throughout the entire utterance rather than being concentrated at the onset. Notably, the last-30-second condition slightly outperformed the first-30-second condition, implying that later segments of the response may contain more stable or representative prosodic evidence of proficiency, potentially due to speakers settling into a more consistent speaking rhythm after the initial articulation phase.

Segment-Based Aggregation and Error Patterns. We further analyzed performance using a segment-based aggregation approach with WhisperX alignment. Each recording was divided into segments, and segment-level scores were aggregated using either the mean or the median. Both strategies achieved performance comparable to the full-length Whisper model, while the median aggregation proved more robust to local inconsistencies and noisy or disfluent segments. This suggests that outlier-prone stretches of speech disproportionately affect global predictions when treated as a single unit, and that segment-wise aggregation can stabilize scoring by emphasizing the speaker's typical performance rather than transient fluctuations.

Error patterns revealed by the confusion matrices (Figure 2) further highlight these differences. With the mean strategy, many level-4 responses were misclassified as level 3, and most level-5 responses were reduced to level 4. Due to the limited number of level-2 samples, the model struggled to classify them correctly. In contrast, the median strategy pro-



(a) Mean—Seen (b) Mean—Unseen (c) Median—Seen (d) Median—Unseen

Figure 2: Confusion matrices comparing mean and median aggregation strategies for proficiency prediction: (a) mean—seen prompts, (b) mean—unseen prompts, (c) median—seen prompts, and (d) median—unseen prompts.

duced more concentrated predictions across both the known and unknown content test sets. Notably, for the unknown content condition, the median strategy yielded more correct classifications for level-5 responses compared to the mean strategy, indicating improved generalization on higher-proficiency learners.

5 Conclusion and Future Work

In this paper, we introduced a segment-based ASA framework for long read-aloud scoring, which addresses the data sparsity and temporal modeling challenges of full-length utterances. Using WhisperX for time-aligned segmentation and weak segment-level labeling, the framework improves supervision granularity and stabilizes the learning of proficiency-relevant speech cues. Experiments on the GEPT corpus showed consistent gains over text-only and speech-only baselines, and revealed that segmentation combined with median aggregation enhances robustness against disfluent or noisy segments. The analysis further highlights that full-length coverage remains essential for reliable scoring, as proficiency cues accumulate beyond early articulation.

Despite these promising results, the framework still assumes that all spoken content aligns with the target passage, whereas learners may occasionally insert off-topic or paraphrastic segments. Since WhisperX already provides high-resolution temporal alignment, future work could exploit this timing information to detect lexical or prosodic deviations from the reference passage, enabling segment-wise content validation rather than treating misalignment as uniform noise. This direction would further extend the framework from holistic scoring toward diagnostic assessment, and could generalize to open-response scenarios where content is not predetermined. Ultimately, incorporating alignment-based semantic verification would improve both the interpretability and applicability of ASA systems in real-world learner-centered settings.

Acknowledgments

This work was supported by the Language Training and Testing Center (LTTC), Taiwan. Any findings and implications in the paper do not necessarily reflect those of the sponsor.

References

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. Cefr-based sentence difficulty annotation and assessment. *arXiv preprint arXiv:2210.11766*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

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

Stefano Bannò and Marco Matassoni. 2023. Proficiency assessment of l2 spoken english using wav2vec 2.0. In *Proceedings of SLT, 2022*, pages 1088–1095.

Lei Chen, Keelan Evanini, and Xie Sun. 2010. Assessment of non-native speech using vowel space characteristics. In *Proceedings of SLT, 2010*, pages 139–144.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018. End-to-end neural network based automated speech scoring. In *Proceedings of ICASSP, 2018*, pages 6234–6238. IEEE.

Eduardo Coutinho, Florian Hönig, Yue Zhang, Simone Hantke, Anton Batliner, Elmar Nöth, and Björn Schuller. 2016. Assessing the prosody of non-native speakers of English: Measures and feature sets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1328–1332, Portorož, Slovenia. European Language Resources Association (ELRA).

Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis. 2020. Investigating the effect of auxiliary objectives for the automated grading of learner english speech transcriptions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2258–2269.

Catia Cucchiarini, Helmer Strik, and Louis Boves. 1998. Quantitative assessment of second language learners' fluency: an automatic approach. In *Proceedings of ICSLP, 1998*, pages paper–0752.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

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 transactions on audio, speech, and language processing*, 29:3451–3460.

Tien-Hong Lo, Fu-An Chao, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen. 2024. An effective automated speaking assessment approach to mitigating data scarcity and imbalanced distribution. *arXiv preprint arXiv:2404.07575*.

Simon W McKnight, Arda Civelekoglu, Mark Gales, Stefano Bannò, Adian Liusie, and Katherine M Knill. 2023. Automatic assessment of conversational speaking tests. In *Proceedings of SLaTE, 2023*, pages 99–103.

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. 2019. Automated essay scoring with discourse-aware neural models. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 484–493.

Yao Qian, Patrick Lange, Keelan Evanini, Robert Pugh, Rutuja Ubale, Matthew Mulholland, and Xinhao Wang. 2019. Neural approaches to automated speech scoring of monologue and dialogue responses. In *Proceedings of ICASSP, 2019*, pages 8112–8116.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chung-Wen Wu and Berlin Chen. 2024. Optimizing Automatic Speech Assessment: W-RankSim Regularization and Hybrid Feature Fusion Strategies. In *Proceedings of Interspeech, 2024*, pages 4004–4008.

Tzu-I Wu, Tien-Hong Lo, Fu-An Chao, Yao-Ting Sung, and Berlin Chen. 2022. A preliminary study on automated speaking assessment of English as a second language (ESL) students. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 174–183, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Tzu-I Wu, Tien-Hong Lo, Fu-An Chao, Yao-Ting Sung, and Berlin Chen. 2023. Effective neural modeling leveraging readability features for automated essay scoring. In *Proceedings of SLaTE, 2023*, pages 81–85.

Exploring the Feasibility of Large Language Model- and Rubric-Based Automatic Assessment of Elementary Students' Book Summaries
大型語言模型結合評分規準於國小學生書籍摘要自動批改之可行性研究

黃琦溱 Qi-Zhen Huang
國立臺灣科技大學數位學習與教育研究所
amity19991122@gmail.com

曾厚強 Hou-Chiang Tseng
國立臺灣科技大學數位學習與教育研究所
tsenghc@mail.ntust.edu.tw

宋曜廷 Yao-Ting Sung
國立臺灣師範大學教育心理與輔導學系
sungtc@ntnu.edu.tw

摘要

摘要寫作為閱讀與寫作整合的高層次語文任務，不僅可評量學生的文本理解能力，也能促進語言表達與重述能力的培養。過去自動摘要批改系統多依賴關鍵詞比對或語義重疊等「由下而上」的方法，較難以全面評估學生的理解深度與文本重述能力，且中文摘要寫作批改研究雖有，但相較於英文仍相對不足，形成研究缺口。隨著大型語言模型（Large Language Models, LLMs）的發展，其在語意理解與生成能力上的突破，為自動摘要批改與回饋帶來新契機。有鑑於此，本研究旨以由上而下的方式探討結合 LLMs 與閱讀摘要評分規準（Rubrics）對學生閱讀摘要批改與回饋之應用潛力，進一步而言，在考量教學資料隱私的情況下，本研究採用 Meta-Llama-3.1-70B 生成電腦摘要，並依據專家所制定的摘要評分規準，其評分涵蓋：理解與準確性、組織結構、簡潔性、語言表達與文法及重述能力五大構面，對學生閱讀摘要進行自動評分與回饋。研究結果顯示，Meta-Llama-3.1-70B 能提供具體、清晰的即時回饋，不僅能指出摘要中遺漏的關鍵概念，也能針對結構安排與語法錯誤提出修正建議，協助學生快速掌握摘要改進方向；然而回饋多偏向表面語言與結構調整，在語言表達、修辭多樣性及重述能力等高層次語文能力評估上仍存在限制。整體而言，LLMs 可作為形成性評量與教學輔助工具，提升評分效率，但需結合教師專業判斷與回饋以補足深層概念與策略性寫作指導，促進學生摘要寫作能力的發展。

關鍵字：大型語言模型、評分規準、摘要自動批改、自動評分

1 緒論

摘要寫作為一整合閱讀與寫作的高層次語文任務，其核心目的在於針對文本的主旨與次要資訊進行簡化與重組，透過精練語言傳達關鍵思想，並展現對文本情感、觀點與結構設計的理解（Özdemir, 2018）。此過程涉及訊息擷取、意義建構與內容統整，能有效反映讀者對文本的理解深度，因而常被用作評量閱讀理解能力的重要指標（Özdemir, 2018）。此外，將冗長內容濃縮為言簡意賅的摘要，亦對學生的文字組織與語言表達能力構成挑戰（Li, J. & Wang, Q, 2021），顯示此任務不僅評估閱讀理解能力，更是一項寫作能力的考驗（Chew et al., 2019；Nelson & King, 2022）。

摘要寫作結合理解與表達歷程，透過適當的摘要策略教學與訓練，能提升學生的記憶保留與理解表現（Sung et al., 2016），並促進對文本訊息的掌握與長期知識鞏固（Graham & Harris, 2000；Silva & Limongi, 2019）。此外，學生在摘要中所呈現的內容準確性與組織結構，與其閱讀理解能力高度相關（Kintsch & van Dijk, 1978；Perfetti, 1985）。藉由學生作品，教師亦可辨識學生的理解盲點與認知偏差（Casteel & Isom, 1994），從而作為調整教學的重要依據。摘要評量的實施，亦為教師提供一項具操作性與客觀性的閱讀理解檢核工具（Afflerbach et al., 2008），可靈活應用於不同學科與文體，並支持後設認知能力的培養（Pressley & Afflerbach, 1995）。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 153–166
November 20-22, 2025 ©2025 Association for Computational Linguistics

儘管摘要寫作具備高度教學價值與評量功能，但在實務操作上，教師卻常因批改負擔過重而減少此類任務的指派（Sung et al., 2016）。摘要批改不僅耗時，亦難以提供即時且個別化的回饋，限制了學生學習與改進的機會（Cheng et al., 2018）。同時，傳統摘要測驗多仰賴人工評分，缺乏標準答案導致評分主觀性高，若批改人員沒有經過訓練，則有可能發生評分者一致性不高的情況，將影響評量的信效度與公正性（Mathews, 1985；Bachman, 1990；Chen et al., 2019）。一個能力的培養需要多次的練習，因此，如何增加練習的次數並提升摘要批改與回饋的品質成為當前教學改革的重要課題。

為解決人工批改的困境，研究者開始運用自然語言處理（Natural Language Processing, NLP）與機器學習（Machine Learning, ML）技術，開發自動摘要評分與回饋系統。這類系統強調能自動化並即時分析學生摘要內容並提供個別形成性回饋，降低教師負擔並提升學學生提交作業的積極度(Nieminen & Isohanni, 2020) 與學習效率 (Sung et al., 2016；Cheng et al., 2018)。例如，潛在語義分析（Latent Semantic Analysis, LSA）便常被用來計算學生摘要與標準摘要的語意相似度，評估是否涵蓋關鍵概念（Kintsch et al., 2000；Wade-Stein & Kintsch, 2004；Landauer et al., 2009）。

然而，過去的研究大多仰賴大量人工標註語料 (Woods et al., 2017)，以專家撰寫的標準答案或電腦摘要來作為摘要自動批改系統評估學生摘要品質好壞的依據（ Kintsch et al., 2000；Wade-Stein & Kintsch, 2004；Landauer et al., 2009）。這樣的方式雖可建立可靠的比對基準，但整體過程往往需投入大量時間與人力成本，標註與比對工作繁瑣且效率低下，使得系統在大規模應用與長期維護上面臨挑戰。且有研究指出，若回饋資訊過於龐雜、模糊、不具體，反而可能增加學生認知負荷，降低學習動機與自我調節能力（Sung et al., 2016；Kim et al., 2021）。此將可能導致學生在自我調節學習上成效有所差異，難以達成因材施教（Sung et al., 2016；Chew et al., 2019）。若為因應跨領域或跨語言使用，則須重新調整系統模型參數與語料庫，此調整將必然增加維護成本 (Lagakis & Demetriadis, 2021) 。

回顧過去自動批改技術的發展，大多屬於「由下而上」（bottom-up）的設計邏輯，主要聚焦於單一語言並透過技術來將文字進行量化，以符合特定摘要評分指標的需求。進一步而言，此類方法多依賴特徵工程（feature engineering）與可量化語言指標（如句長、關鍵詞出現率、語意相似度等）來評估學生摘要品質（Landauer et al., 2009；Kintsch et al., 2000；Woods et al., 2017），而非從整體語義脈絡或語用功能進行整體性理解。如：Landauer 等人（2009）利用潛在語義分析（Latent Semantic Analysis, LSA）來比對學生摘要與標準答案的語意相似度，以判斷是否涵蓋核心概念；Kintsch 等人（2000）亦在「Summary Street」系統中應用 LSA，即時計算學生摘要與原文的語意相似度，協助學習者改善摘要品質。這些方法雖能展現一定程度的自動化與客觀性，但其依賴標準答案與語料庫的特性，限制了對文本結構變異性與語意重組能力的評估準確性（Sung et al., 2016）。此外，當面對需跨主題、跨文本應用或須結合修辭判斷與語境理解的寫作任務時，其效能亦顯不足。例如，Partanen 等人（2018）指出，雖然主流自動評分模型能有效檢測句法正確性與表層流暢度，但卻難以辨識文本在篇章層次特徵（discourse-level features）上的銜接與整體連貫性。綜上所述，未來自動批改系統的設計需更具整體語意理解與彈性應對能力，以突破傳統「由下而上」模型的侷限（Dikli, 2006）。

鑒於傳統自動評分系統多採「由下而上」的設計邏輯，難以應對語義整合與跨文本應用的挑戰。近年來，大型語言模型（Large Language Models, LLMs）之發展為此領域帶來新契機。LLMs 具備優越的語意理解與生成能力，能從整體文本層次進行語義評估與改寫建議，展現「由上而下」（top-down）整體理解的評分潛能（Cheng et al., 2018；Botarleanu et al., 2022）。例如，Morris 等人（2024）採用 Longformer 這類 LLM 進行學習者於智慧教科書撰寫的結尾摘要評估。研究透過微調（fine-tuning）模型，使其依據摘要內容與原始文本，在結構與內容兩方面提供即時且具體的回饋。結果顯示 LLM 對於摘要品質的評估具有顯著準確性，也驗證了 LLM 在摘要自動批改上的實務可行性與效果。

大型語言模型（LLMs）在自然語言處理領域的突破，尤以其深層語意理解與語境生成能力著稱，為評分與教學提供全新視角。與傳統模型不同，LLMs 不需仰賴大量人工設計的特徵，而是透過深度學習架構，例如：Transformer，掌握語言的統計規律與語義脈絡（Vaswani et al., 2017；Brown et al., 2020），能有效辨識文本中的主旨、層次結構與語用功能，進行更貼近人類的整體性評估與改寫建議。基於此優勢，LLMs 極有潛力在統一評分規準的基礎上，發展出兼具教學診斷功能與自動批改效能的系統。例如，教師可使用同一份評分規準，作為指引學生摘要寫作的教學工具，亦可導入 LLMs 進行即時批改與回饋，實現「教學—評量—修訂」三者整合的目標（Sung et al., 2016；Chew et al., 2019；Botarleanu et al., 2022）。此一設計理念不僅可強化學生語意整合與自我修訂的能力，也可提高評分規準在實務操作上的一致性與可遷移性。因此，本研究嘗試以 LLMs 為基礎，發展一套整合評分與回饋的智慧摘要批改系統，冀能提升教學現場的評量效能並促進學生摘要寫作能力之長期養成。

本研究旨在探討結合大型語言模型（Large Language Models, LLMs）與閱讀摘要評分規準，是否對書籍閱讀摘要的批改回饋能有良好的準確性及回饋品質。過去研究指出，教師在批改學生摘要時常面臨主觀性高與耗時的困境（Sung et al., 2016），而目前的自動批改系統多聚焦於格式檢核與詞彙密度等表層特徵，難以全面審視學生對文本的理解深度與重述能力（Zhang & Litman, 2015）。因此，本研究基於「由上而下」之語意理解策略之理論基礎，設計一套整合大型語言模型與評分規準的自動摘要批改與回饋系統來為學生閱讀摘要評分，以提供兼具準確性及形成性價值的自動批改結果，並驗證 LLMs 搭配評分規準運用於閱讀摘要評分與回饋的可行性，並分析其語義層次評估與生成式回饋的品質與限制，期能結合人類評分的語用敏感性與自動化系統的一致性與即時性，應用於教學現場，作為形成性評量與學習歷程回饋工具。進一步提升評量效能並支持學生摘要寫作能力的長期培養（Sung et al., 2016；Lu, 2011）。

2 文獻探討

早期自動評分系統多以潛在語義分析（Latent Semantic Analysis, LSA）為基礎，其核心技術透過詞彙-文件矩陣捕捉文本之間的語義相似度（Landauer et al., 2009），用以比較學生摘要與原文之間的語意重合程度。如「State the Essence」與「Summary Street」(Kintsch et al., 2000）等系統，用於協助學生進行摘要訓練並提供即時計算摘要與原文的相關性，協助學生改善摘要寫作表現，（Sung et al., 2016）。結果顯示該系統能顯著提升學生的寫作表現（Wade-Stein & Kintsch, 2004）。然而，此類技術著重於語義重疊，對文本的邏輯結構與語言表現仍缺乏細緻辨識能力，回饋也較為制式，限制其在高層次寫作評量的應用效益（Chew et al., 2019）。此外，研究者亦常使用主成分分析（Principal Components Analysis, PCA）將多個評分項目整合為如「內容（Content）」與「措辭（Wording）」等關鍵構面，以提升評分解釋力與一致性（Lu, 2011；Chen et al., 2019）。

除語義相似度外，亦有研究著重於文本層面的語言特徵分析，發展出以剖析樹（Parsing Trees）與自然語言處理指標為基礎的評分方法，以評估學生的詞彙選擇、語法結構運用以及篇章組織能力等寫作品質。例如，Coh-Metrix 能自動偵測文本中的複雜句型、詞彙多樣性、語法正確性與篇章連貫性，進一步評估學生語言使用的成熟度（Graesser et al., 2004；Burstein et al., 2013）。此類技術不僅提升了評分的細緻度，亦有助於捕捉文本內部結構與語言風格，提升評量的準確性與解釋力。隨著機器學習技術的成熟，近年亦有研究嘗試運用如類神經網路（Artificial Neural Networks, ANN）、支援向量機（Support Vector Machines, SVM）(Cortes & Vapnik, 1995）…等等監督式學習（supervised learning）方法訓練出評分模型以提升預測準確率（Cheng et al., 2018；Zhang et al., 2020）。這些方法強調從大量標註資料中學習語言特徵與評分邏輯，使自動化系統更能模擬人類評分者的判斷。整體而言，機器學習模型能有效處理高維度的特徵以展現良好的效能。

近年來，大型語言模型（Large Language Models, LLMs）如 GPT（Brown et al., 2020）、 Longformer（Beltagy et al., 2020）等的出現應用，為自動評分系統帶來突破性發展。LLMs 能處理長文本並具備語意整合與生成能力，能更準確地評估學生摘要與原文之間的語意對應與語言品質（Beltagy et al., 2020）。例如，Botarleanu et al.（2022）指出，大型語言模型 (LLMs) 可結合語意理解與語言生成能力，不僅能進行評分，也能提供近似於真人教師的具體語用回饋及寫作建議，幫助學生進行摘要修訂，顯著提升學習成效。此外，LLMs 可整合不同面向的語言指標，從內容涵蓋、邏輯組織到語體風格進行整合性評估，是目前最具潛力的智慧批改技術。總體而言，技術演進已從早期重視效率的語義相似度評估，邁向結合語意判讀、風格辨識與即時生成回饋的智慧化系統，有望更準確模擬人類評分邏輯，並提升摘要寫作教學與摘要寫作學習效能（Sung et al., 2016；Cheng et al., 2018）

由上述文獻可知自動評分技術雖歷經多階段演進，但目前的研究仍存在幾項關鍵缺口。其一，絕大多數系統與資料集皆為英文語境，中文語境下之摘要寫作特性與語用風格尚缺乏充分探討，且現有自動批改系統多依賴商用 LLMs，如：OpenAI 的 ChatGPT，進行部署，將開發完成的程式或應用程式，從開發環境推送到可以讓使用者真正使用的測試伺服器、生產環境或雲端平台，引發資料隱私、使用成本與模型調整彈性等問題（Wang & Chiang, 2022；Zhao et al., 2023），在教學現場信任度與實務上為一大顧慮（Ding et al., 2023）。其次，多數自動批改系統仍延續傳統「由下而上」(bottom-up）的技術邏輯，如比對關鍵詞（Louis & Nenkova, 2013）、計算句距與句子重疊率（Lin, 2004）、以及套用句型規則（Attali & Burstein, 2006）等方法，強調特徵抽取與評分規則對應，雖具可計量性與一定程度的形式驗證，卻難以處理語意整合與篇章理解等高層次語文歷程（Louis & Nenkova, 2013；Somasundaran et al., 2014）。這樣的限制使得自動批改系統難以提供貼近教學目標的深層回饋，無法協助教師辨識學生在內容統整、語意建構與語體運用上的實質困難。此外，教師在教學上若倚賴此類系統，可能會誤導學生過度追求表面形式正確，而忽略摘要寫作中更核心的思維組織與語言表達能力，進而限制其語文素養的深度發展。

有鑑於此，本研究關注的核心即為 LLMs 在摘要評分、批改任務中所展現之「由上而下」（top-down）語意建構能力。根據語言理解與心理語言學理論，「由上而下」(top-down）的技術邏輯強調讀者、系統先有整體語境與主旨的理解，再回溯文本細節進行詮釋與分析（Rumelhart, 1980；Kintsch, 2004），與語文教學中強調的深度理解與篇章層次建構密切契合（Nelson & King, 2022），模擬教師從「整體—局部—整體」的評分邏輯（Chew et al., 2019；Botarleanu et al., 2022）。相較傳統仰賴詞彙與句法特徵的「由下而上」(bottom-up）技術邏輯，更能貼近教師實際的教學與評分思維，體現語意層次的整合與推論，對學生摘要整體理解的評估更具語用敏感性（Liu et al., 2022）。在此基礎上，本研究的設計構想即是利用 LLMs 的語意理解優勢，先讓模型理解研究所設計之評分規準與電腦摘要，藉此建立一個「整體語境—評分標準—摘要內容」的參照框架，再以此為依據批改與評分學生的書籍摘要。這樣的流程不僅展現了「由上而下」的語意驅動邏輯，也使得模型能夠模擬教師在形成性評量與教學診斷中所採取的整體判斷與脈絡化詮釋，展現更高的評分一致性與教學應用可行性。

3 研究設計

3.1 實驗流程

本研究使用 Meta-Llama-3.1-70B（Meta AI, 2024）作為電腦自動摘要與評分工具，首先透過提示詞(prompt)設計引導模型對書籍產生電腦摘要，並根據電腦摘要與評分規準對學生摘要進行評分與回饋。最後，將電腦摘要、修訂完畢之評分規準、以及學生摘要一併輸入模型中進行自動評分，評分構面涵蓋「理解與準確性」、「組織結構」、「簡潔性」、「語言表達與文法」以及「重述」。結束後以人工進行檢視、探討 Meta-Llama-3.1-70B 模型在摘要評量任務中的可行性與應用潛力。總體實驗流程如下圖 1 所示：

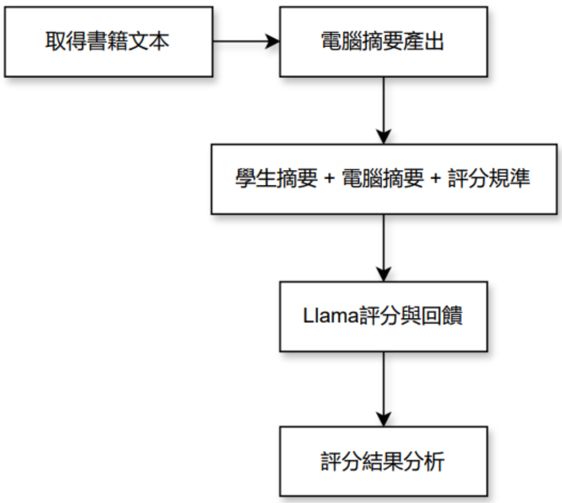


圖 1：基於 Meta-Llama-3.1-70B 產生電腦摘要，並根據此電腦摘要與評分規準對學生摘要進行評分、回饋及結果分析之實驗流程圖

3.2　研究對象與研究工具

3.2.1 書籍文本與學生摘要

本研究與 SmartReading 適性閱讀平台(教育部校園數位內容與教學軟體中心，2007)合作，取得書籍文本及參與者（ 國小學童 ）閱後所謄寫之 200 字摘要。SmartReading 適性閱讀系統為具提供閱讀能力診斷與適性圖書推薦能力的智慧學習平台，學生閱讀之書籍皆於接受中文適性閱讀能力診斷（Diagnostic Assessment of Chinese Competence，DACC）後，由系統判斷中文閱讀程度，再從文本可讀性指標自動化分析系統（ Chinese Readability Index Explorer，CRIE)，依參與者程度推薦之書單內書籍。本研究選定國小人文領域之書籍，內容是一本闡述關於戰爭所引發貧窮與飢餓的書；並針對平台參與者閱讀此一書籍後所撰寫之摘要進行亂數挑選，共計 30 篇。

3.2.2 大型語言模型 Meta-Llama-3.1-70B 與電腦摘要

本研究使用 Meta-Llama-3.1-70B（Meta AI, 2024）作為電腦自動摘要與評分工具，透過提示語(prompt)設計引導模型對書籍產生電腦摘要，並對學生撰寫之摘要根據電腦摘要與評分規準進行評分與回饋生成。該模型基於 Transformer 架構，結合自注意力（Self Attention）技術，能高效捕捉語境細節，提升語言理解、長文本處理、分析與邏輯推理能力。

且過去研究發現大型語言模型生成的摘要展現出更佳的事實一致性 (factual consistency)，且較能避免外在幻覺(extrinsic hallucinations) (Pu et al., 2023)，在流暢性 (fluency) 、連貫性(coherence)上亦表現良好，能靈活調整輸出文長，並全面涵蓋文本內容主旨大意 (Pu et al., 2023)。基於其品質與有效性，本研究將 Meta-Llama-3.1-70B 所生成之電腦摘要為研究工具。

3.2.3 閱讀摘要評分規準

本研究為客觀且系統性為學生所撰寫之閱讀摘要評分，設計一閱讀摘要評分規準。該規準乃參考相關文獻(Chen et al., 2023；Morris et al., 2024；Özdemir, 2018)之摘要寫作與評量標準並結合摘要寫作教學核心能力所設計，最後經閱讀寫作專家審視、共同修訂完成。該閱讀摘要評分規準最終涵蓋五大評分構面，包括：(一)理解與準確性、(二)組織結構與邏輯條理、(三)簡潔性、(四)語言表達與文法、以及(五)重述能力。各構面包含一至數個子面向，每一面向皆設計五點等級，由「優」(5 分) 至「不足」(1 分)，旨在檢視學生摘要在內容、結構、語言與重述等多重層面的整體品質。評分規準示意表如表 1：

表 1：閱讀摘要評分規準示意

評分構面	子面向	概念型定義
理解與準確性	核心概念	是否精確呈現原文主旨、核心概念
	重要事件	是否提取原文關鍵資訊、重要事件
	理解正確	是否對提取出的關鍵資訊、重要事件有正確的理解和詮釋

組織結構與邏輯條理	結構層次與完整性	結構清晰、層次分明，有明顯段落分工，整體架構完整
	邏輯與條理性	資訊間邏輯關係明確、前後一致，內容具因果、對比、遞進等關係，條理分明
	句段連貫性	句、段間銜接自然，整體內文承接連貫流暢
簡潔性	精煉	語句是否精簡、準確
	冗贅	是否含重複、無關或多餘內容
語言表達與文法	用字精確	詞語選擇是否精確、貼切
	修辭	是否有適當語言技巧或表達效果
	語句流暢性	語序是否自然、易讀
重述	重述	是否能以自身語言轉述原文、展現語言轉換能力

首先，理解與準確性著重檢視學生是否能精確掌握原文主旨、核心概念與重要事件，並正確理解所摘取的資訊。其次，組織結構與邏輯條理關注摘要內部之篇章架構、段落層次、邏輯連貫性及銜接流暢度。第三，簡潔性評估語句是否能夠精煉傳達核心資訊，並避免冗贅或重複的描述。第四，語言表達與文法評估學生在詞語選擇、修辭手法、語句流暢性與文法正確性等方面的表現。最後，重述能力則檢視學生是否能以自身語言轉述原文，展現語言轉換與重組能力，而非僅依賴原文句構或直接抄錄。

此一評分規準的建構，旨在兼顧摘要寫作任務中「理解—組織—表達」的多重層面，並能兼具效度與信度。規準經由閱讀專家審視與修訂，確保其評分標準具備操作性與一致性，適用於本研究對學生閱讀摘要作品的系統性分析。透過該規準，本研究得以檢驗學生在不同面向的摘要表現，並進一步審視以真人角度來看電腦自動評分結果的優缺。

4 研究結果與分析

4.1 電腦自動評分結果與分析

本研究共蒐集 30 篇學生摘要，經去識別化，移除個人可識別資料後，使資料無法再直接或間接連結到特定學生，以保護其隱私，再由 Meta-Llama-3.1-70B 根據五大構面進行自動評分，並計算描述性統計與總體分數，以呈現模型在各構面上的評分。

表2：Meta-Llama-3.1-70B根據電腦摘要及評分規準針對學生摘要評分結果

評分構面		*M*	*SD*	最高分	最低分
理解與準確性	(n = 30)	3.72	.81	5	2
組織結構	(n = 30)	3.45	.95	5	1
簡潔性	(n = 30)	3.10	.88	5	1
語言表達與文法	(n = 30)	3.60	.77	5	2
重述	(n = 30)	2.95	.92	5	1
總體分數		3.36	.84		

從整體分數來看，模型在五構面（理解與準確性、組織結構、簡潔性、語言表達與文法、重述能力）的平均分數介於 2.95 至 3.72 間，總體平均分數為 3.36 分（*SD* = .84)，顯示學生的表現大致落在中上水準，但不同構面間仍有所差異，且在部分高層次語文能力的辨識上仍具限制。

在「理解與準確性」方面，(*M*=3.72, *SD*=.81)，學生摘要表現平均分最高，僅少部分學生出現偏離核心概念或只摘錄細節的情況。在此構面模型能準確評估學生是否涵蓋原文的核心概念，並能辨識主旨是否遭到偏移或僅呈現瑣碎細節；**若摘要能精簡並正確提取出原文的核心概念，則獲得較高分數，若無，則分數就會顯得較低**。表 3 為批改結果與人工觀察之結果。

表 3：「理解與準確性」批改結果示意

學生摘要	電腦自動批改結果	觀察與分析
世界上有一群人，過著沒有錢，吃不飽，穿不暖的生活，沒有舒適的家，食物或乾淨的水，人很容易生病。誰都不	核心概念：2 分（待加強）：學生摘要未能完全掌握原文的核心概念，例如貧窮的定義、原因和影響等。	電腦摘要指出：貧窮的意思就是擁有很少的錢，甚至完全沒有錢。背後的原因可能是氣候、是戰爭，而這些

希望因為貧窮就受到不同的態度對待。知曉世界事，同理他人處境，擴展國際觀，培養成為未來領袖的胸懷。這書內容符合教育部頒布的中小學國際能力指標，以深入淺出的文字及精美插圖，將遙遠的國際事件與孩子的生活經驗緊密扣合，幫助他們理解發生在世界各地的重大議題，分析這些事件背後的原因，以及未來可能的解決方式。不管年紀多小，都可以對這世界多一點理解、多一點關心、多一點出，就從現在一起閱讀開始。		都可能造成貧窮、飢餓、以及工作、健康方面等生活上的影響。批改結果準確：學生只提到此書有提及貧窮的定義、原因和影響，但沒有具體列舉。
	重要事件：2分（待加強）：學生摘要未能提取出原文的重要事件，例如戰爭、氣候變化等。	電腦摘要指出：戰爭也會造成貧窮與飢餓。批改結果準確：學生只提到國際事件、重大議題，但並未指出戰爭。
	理解正確：2分（待加強）：學生摘要對原文的理解存在誤差，例如將貧窮等同於沒有錢、吃不飽、穿不暖等。	電腦摘要指出：貧窮的意思就是擁有很少的錢，甚至完全沒有錢。批改結果準確：貧窮可能只是擁有「很少」的錢，並非一定「沒有」錢。

在「組織結構」構面中，(M=3.45, SD=.95)，分數離散程度較大，部分學生能清晰展現段落邏輯，**但若使用隱性推論或非明示銜接策略時，模型無法精準判斷其合理性，導致部分評分偏低**。批改結果與人工觀察之結果如表 4 所示。

表 4：「組織結構」批改結果示意

學生摘要	電腦自動批改結果	觀察與分析
這本書的主角是貧窮與飢餓的人。九個人就有一個人沒有足夠的食物吃。造成貧窮與飢餓的有戰爭、天災、工作的地方關閉和身心障礙。貧窮的人很容易生病，因為他們沒有乾淨的水和食物。貧窮的小孩因為太飢餓根本無法學習，長大後還是會變成貧窮和飢餓的生活。慈善組織是助人的團體，他會教他們種田，協助建照水槽以及廁所，他們還會蓋學校和醫院來幫助貧窮的人。助人為快樂之本。	結構層次與完整性：2分（待加強：段落分工稍顯模糊，各段主旨多有重疊或資訊羅列順序不當，整體篇章結構鬆散）	批改結果準確：學生摘要並未分段，學生先提出貧窮的影響(容易生病)，再提出造成貧窮與飢餓的是因為沒有乾淨的水和食物？後又繼續說明貧窮與飢餓的影響（無法學習，長大後還是會變成貧窮和飢餓的生活)，架構、邏輯脈絡稍顯模糊不清。
	邏輯與條理性：2分（待加強：明顯邏輯混亂，條理順序錯置，無法清楚呈現資訊之因果、對比、遞進等關係）	
	句段連貫性：2分（待加強：缺乏關係連接詞，句段間顯得斷裂）	批改結果準確：例如，學生摘要在提到慈善組織之前，沒有轉承的連接詞，最後說明完慈善組織時，又直接引用名言佳句作為單句心得，整體句子間之脈絡結構顯得斷裂。

在「簡潔性」的評分上，(M=3.10, SD=.88)，**模型展現出對冗贅與重複語句的高度敏感度，能夠檢測學生是否過度重複同一概念，並建議刪減或合併。然而，其對「必要補充細節」與「冗餘內容」的區辨能力仍顯不足，造成評分偏低**。例如，若學生為避

免誤解而加入額外解釋，模型可能會將其視為不必要的贅詞並扣分。批改結果示意如表 5：

表 5：「簡潔性」批改結果示意

學生摘要	電腦自動批改結果	觀察與分析
這本書介紹了世界上有平窮的家庭和有錢的家庭，平窮的家庭過的日子和有錢的家庭過的日子。這本書一開始介紹了，平窮的家庭和有錢的家庭住的地方。之後還介紹了，平窮的家庭和有錢的家庭學習的方式有錢的人是會比較活潑平窮的人恵英未被霸令或必較虛弱。最後介紹了，有錢的家庭會怎麼讓平窮的家庭對他們的人身感到快樂。	精煉：1 分（不足）：學生摘要的語句未能精煉，未能有效地濃縮原文內容。	批改結果準確：學生摘要可省略後方重複的語句，如：「平窮的家庭和有錢的家庭」，學習方式也並非此書摘要重點，應可省略並補足核心事件，因此學生摘要著實未能有效地精確濃縮原文內容。
	冗贅：1 分（不足）：學生摘要含有重複、無關或多餘的內容，影響了摘要的焦點。	批改結果準確：學生摘要一直反覆出現「平窮的家庭和有錢的家庭」，重複的語句內容佔據很高比例的篇幅，會讓人有以為學生重複繕打相同內容的錯覺。

「語言表達與文法」方面，(M=3.60, SD=.77)，是學生摘要表現平均分數僅次於理解與準確性的構面；雖**模型對顯而易見的文法錯誤（如：主謂不一致或結構不完整）具有良好的判斷，並能指出需要修正的句子。然而，在修辭多樣性、語氣適切性與表達自然度等高層次語言層面，模型的判斷仍相對粗略**。批改結果示意如表 6：

表 6：「語言表達與文法」批改結果示意

學生摘要	電腦自動批改結果	觀察與分析
有些孩子買不起很好的房子，但有些人卻買得起，全世界至少有四分之三的人沒錢買食物和房子很乾淨的水，有些人則是自己種東西來吃。戰爭會讓在種田的人無法繼續中還會有人故意埋設炸彈在田裡讓在種田的人無法繼續中。	用字精確：2 分（待加強：學生摘要未能準確傳達原文核心意思，用字遣詞不夠精確。）	批改結果準確：學生摘要後半段「戰爭會讓在種田的人無法繼續中還會有人故意埋設炸彈在田裡讓在種田的人無法繼續中。」就出現了錯字，「中」應修正為「種」。
	修辭：1 分（不足：學生摘要未能運用適當的語言技巧，整體表達生硬、乏味。）	批改結果準確：學生摘要語句多維平鋪直述，幾乎未見修辭技巧。
	語句流暢性：2 分（待加強：學生摘要的句構不夠自然流暢，語序不當。）	批改結果準確：學生摘要句構著實仍有進步空間，如摘要後半段「戰爭會讓在種田的人無法繼續中還會有人故意埋設炸彈在田裡讓在種田的人無法繼續中。」中間無斷句，前後轉承也不流暢。

最後，在「重述能力」方面，(M=2.95, SD=.92)，為學生摘要表現平均分最低的構面；**模型雖能偵測直接抄寫原文或僅作同義詞替換的情形，但對於部分改寫（如：句式重組）與完整重述之間的細緻差異仍顯不足，導致在此構面的判斷偏向保守**。批改結果示意如表 7：

表 7：「重述能力」批改結果示意

學生摘要	電腦自動批改結果	觀察與分析
這本書告訴我們什麼人正在忍受貧窮與飢餓, 什麼事會造成貧窮與飢餓. 戰爭付出的代價只會讓國家更貧窮, 人民生活得更可憐還要被迫逃離自己的家園變成難民. 天災也是造成貧窮的原因之一, 天災包含了地球暖化, 乾旱, 颱風,水災. 還有貧窮造成的健康問題,例如沒有乾淨的水源和沒有足夠的食物的吃這些都需要有能力有愛心的人伸出援手. 慈善組織的幫助人的團體, 他們幫助有需要的家庭和孩子. 他們也經營食物銀行給貧窮的人. 他們也會幫助人們改善生活比如提供工具給貧窮的人蓋房子, 提供船給貧窮的漁村並幫助她們開設商店或開創其他事業. 助人為快樂之本, 很多事情你也做得到！	幾乎完全未重述，內容多為直接抄寫或機械性重組原文句子，與原文文字敘述高度重疊。	批改結果準確：學生摘要雖有具體舉出重要事件，但多數語句都是擷取原文而來，如：慈善組織「幫助有需要的家庭和孩子」. 他們也「經營食物銀行」給貧窮的人. 他們也會「幫助人們改善生活」比如「提供工具給貧窮的人蓋房子」，「提供船給貧窮的漁村」並幫助她們「開設商店」或「開創其他事業」。除了轉承詞以外，少有自己重述部分。

5 結論

Llama-3.1-70B 在部分評分構面展現出高度可靠性，特別是在「理解與準確性」與「組織結構」兩個層面。模型能夠準確辨識學生是否涵蓋文本的核心概念，並有效評估摘要中段落之間的邏輯銜接與結構完整性，顯示 Llama-3.1-70B 具備自動化分析文本整體內容與組織的能力。此外，模型能快速處理大量文本並提供即時回饋，對於教師在大班教學或大量寫作作業批改中，具有明顯的效率優勢。這些技術特徵使 Llama-3.1-70B 成為摘要評量與教學輔助的潛在工具，能有效分擔教師的工作負擔，並提升評量過程的即時性與一致性。

然而，Llama-3.1-70B 在部分評分構面上的表現仍具限制。在「語言表達與文法」的評估中，雖然模型能準確偵測文法錯誤，但在修辭多樣性、語氣適切性與表達自然度的判斷上，仍存在限制。同樣地，在「重述能力」的構面上，模型能辨識學生是否直接抄寫原文，但對於「部分改寫」與「完整重述」的細緻區分能力不足，導致其評分結果偏低。這些限制反映出 Llama-3.1-70B 雖能有效掌握摘要的內容與結構，但在語言層次的深層分析與語意重組的敏感度上仍有其限制。

此外，模型在回饋生成上，本研究發現 Llama-3.1-70B 所生成的建議通常結構清晰且具體，能夠有效指出學生在語言表達與組織上的不足，並提供明確的修正方向。然而，其生成的建議傾向於聚焦表層語言與結構的修改，缺乏針對深層理解與批判思維的引導。從實際教學應用的角度來看，Llama-3.1-70B 雖能提供即時且具體的回饋，協助學生改善語言表達與組織結構，其侷限性仍需正視。若學生過度依賴模型回饋，可能僅停留於表

層修訂，忽略對文本意涵的深層理解與批判性思維的養成。

6 研究限制與未來發展

本研究樣本限於小學，且研究文本僅涵蓋一類型，未涵蓋多樣化體裁，因此研究結果在其他寫作情境下的適用性仍待驗證。其次，本研究僅使用單一大型語言模型進行測試，未與其他 LLMs 進行比較，未比較不同 LLMs 的效能，無法全面呈現不同開源大型語言模型在摘要評量與回饋上的差異與相對優劣。這些因素皆可能影響研究結果的廣泛適用性，未來仍需更多實證研究來驗證其在不同教學情境中的可行性。

基於上述發現，本研究提出以下應用與研究建議。首先，在教學現場，教師可將 Llama-3.1-70B 視為輔助工具，運用其快速診斷與初步修訂建議的功能，幫助學生即時修正語言與結構上的問題，然而教師仍需補充深層的概念性指導與批判性思考訓練，以避免學生僅停留於表層學習，確保學生能同時兼顧語言表達與內容理解。其次，對學生而言，模型回饋應被視為修訂的參考依據，而非最終標準，並透過自我反思與反覆修訂，逐步培養更高層次的摘要能力與自主學習意識。再者，於系統設計層面，未來可針對 Llama-3.1-70B 的修辭敏感度與重述能力進行優化，並發展更具互動性的回饋形式，如：提供範例對照或逐步引導，以增進其在寫作教學中的輔助效果。最後，後續研究可擴展樣本規模、引入多樣化文本類型，並比較不同模型的表現，從而更全面檢視 Llama-3.1-70B 在摘要評量與寫作教學中的應用價值。

基於上述限制，未來研究可從數個方向進一步拓展。其一，可擴展樣本來源與規模，以檢視不同背景學生在使用模型輔助下的學習成效差異。其二，可引入不同文體與多樣化文本類型，評估模型在不同寫作任務中的表現。其三，可比較不同 LLMs 之間以及 Llama-3.1-70B 與專家評分的相關性，進一步分析其評分準確度與回饋品質的差異。最後，未來研究可探索設計更智慧化的學習系統，結合模型的即時回饋與教師的深層指導，發展出能動態調整回饋深度與內容的機制，以更有效地支持學生的個別化學習需求。

綜言之，本研究顯示 Llama-3.1-70B 於摘要寫作教學中具有可觀的潛力，尤其在提供即時回饋與輔助評量上能發揮效能。然而，模型仍不足以完全取代專家評分與指導，教師的專業判斷與深層引導仍為不可或缺的部分。若能在教學現場中妥善整合 Llama-3.1-70B、教師專業以及專家回饋，將有助於建立更完善的「寫作—回饋—修訂」循環，從而促進學生的寫作發展與自主學習能力。兩者若能結合，將有助於建構多層次的教學支持體系。

致謝

本研究承國科會研究計畫 114-2628-H-011 -002 -MY3、國立臺灣科技大學教育部高教深耕計畫特色領域技職賦能研究中心及國立臺灣師範大學教育部高教深耕計畫華語文科技中心補助。

謹此致謝

文獻

Adeshola, Ibrahim & Adepoju, Adeola. (2023). The opportunities and challenges of ChatGPT in education. Interactive Learning Environments. 32. 1-14. https://doi.org/10.1080/10494820.2023.2253858.

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher*, *61*(5), 364–373. https://doi.org/10.1598/RT.61.5.1

Attali, Y., & Burstein, J. (2006). Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. Journal of Technology, Learning, and Assessment, 4(3). *Journal of Technology, Learning, and Assessment, 4.*

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*. https://arxiv.org/abs/2004.05150

Botarleanu, S. M., Henschel, A., Hämäläinen, S., & Al-Sabbagh, M. (2022). Can large language models provide feedback to student writing? *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*, 648–652.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 55–67). Routledge.

Casteel, C. P., & Isom, B. A. (1994). Reciprocal teaching of comprehension strategies with students with learning disabilities. *Learning Disability Quarterly, 17*(2), 169–184. https://doi.org/10.1086/461828

Chen, C., Li, Z., Peng, Z., & Li, Q. (2023). *ALens: An adaptive domain-oriented abstract writing training tool for novice researchers*. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). ACM. https://doi.org/10.1145/3544548.3581512

Chen, M., & Zheng, Y. (2022). Book review: The Routledge Handbook of Second Language Acquisition and Writing. *Journal of Writing Research*, *14*(2), 287-292. https://doi.org/10.17239/jowr-2022.14.02.05

Chen, Q. (2025). Students' Perceptions of AI-Powered Feedback in English Writing: Benefits and Challenges in Higher Education. *International Journal of Changes in Education*. https://doi.org/10.47852/bonviewIJCE52025580

Cheng, Y.-S., Wu, W.-C. V., & Ku, Y.-M. (2018). Exploring the effects of summarization-based reading strategy instruction on EFL learners' reading comprehension. *Interactive Learning Environments, 26*(3), 427–441. https://doi.org/10.1080/10494820.2017.1337035

Chew, C. S., Lin, D. T. A., & Chen, S. (2019). The effects of a theory-based summary writing tool on students' summary writing. *Journal of Computer Assisted Learning, 35*(3), 435–449. https://doi.org/10.1111/jcal.12349

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Danyluk, A., & Buck, S. (2019). Artificial Intelligence Competencies for Data Science Undergraduate Curricula. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 9746-9747. https://doi.org/10.1609/aaai.v33i01.33019746

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*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning and Assessment*, *5*(1). https://ejournals.bc.edu/index.php/jtla/article/view/1640

Ding, Y., Fu, S., & Yang, X. (2023). The application of ChatGPT in education: Opportunities and challenges. *Education and Information Technologies*. Advance online publication. https://doi.org/10.1007/s10639-023-11886-2

Partanen, N., Lim, K., Rießler, M., & Poibeau, T. (2018). Dependency parsing of code-switching data with cross-lingual feature representations. In I. Kallio, J. Laippala, & J. Puskás (Eds.), *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages* (pp. 1–17). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0201

Pu, X., Gao, M., & Wan, X. (2023). *Summarization is (Almost) Dead. arXiv*. https://doi.org/10.48550/ArXiv.2309.09558

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Graham, S., & Harris, K. R. (2000). The role of self-regulation and transcription skills in writing and writing development. *Educational Psychologist*, *35*(1), 3–12. https://doi.org/10.1207/S15326985EP3501_2

Gutierrez, Fernando & Atkinson-Abutridy, John. (2011). Adaptive feedback selection for intelligent tutoring systems. Expert Syst. Appl.. 38. 6146-6152. https://doi.org/10.1016/j.eswa.2010.11.058

Huawei, S., Aryadoust, V. A systematic review of automated writing evaluation systems. *Educ Inf Technol* 28, 771–795 (2023). https://doi.org/10.1007/s10639-022-11200-7

教育部校園數位內容與教學軟體中心. (2007). *SmartReading 適性閱讀*. https://www.sdc.org.tw/product/smartreading%E9%81%A9%E6%80%A7%E9%96%B1%E8%AE%80/

Khoshsima, Hooshang & Tiyar, Forouzan. (2014). The Effect of Summarizing Strategy on Reading Comprehension of Iranian Intermediate EFL Learners. International Journal of Language and Linguistics. 2. 134-139. https://doi.org/10.11648/j.ijll.20140203.11

Kim, J., Yu, S., Detrick, R. *et al*. Exploring students' perspectives on Generative AI-assisted academic writing. *Educ Inf Technol* **30**, 1265–1300 (2025). https://doi.org/10.1007/s10639-024-12878-7

Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. In R. B. Ruddell & N. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 1270–1328). International Reading Association.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Kintsch, W., Steinhart, D., Stahl, G., Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, *8*(2), 87–109. https://doi.org/10.1076/1049-4820(200008)8:2;1-D;FT087

Kobrin, J. L., Deng, H., & Shaw, E. J. (2011). The association between SAT prompt characteristics, response features, and essay scores. *Assessing Writing*, *16*(3), 154-169. https://doi.org/https://doi.org/10.1016/j.asw.2011.01.001

Lagakis, K., & Demetriadis, S. (2021). Adaptive feedback in intelligent tutoring systems: A review of recent advances. *Educational Technology Research and Development*, *69*(6), 3185–3213. https://doi.org/10.1007/s11423-021-10044-7

Lagakis, P., Demetriadis, S. (2022). Automated Essay Feedback Generation in the Learning of Writing: A Review of the Field. In: Auer, M.E., Tsiatsos, T. (eds) New Realities, Mobile Systems and Applications. IMCL

2021. Lecture Notes in Networks and Systems, vol 411. Springer, Cham. https://doi.org/10.1007/978-3-030-96296-8_40

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2009). *Handbook of latent semantic analysis*. Psychology Press.

Leszczyński, P., Charuta, A., Łaziuk, B., Gałązkowski, R., Wejnarski, A., Roszak, M., & Kołodziejczak, B. (2017). Multimedia and interactivity in distance learning of resuscitation guidelines: a randomised controlled trial. *Interactive Learning Environments*, *26*(2), 151–162. https://doi.org/10.1080/10494820.2017.1337035

Li, J., Wang, Q. Development and validation of a rating scale for summarization as an integrated task. *Asian. J. Second. Foreign. Lang. Educ.* **6**, 11 (2021). https://doi.org/10.1186/s40862-021-00113-6

Lin, C.-Y. (2004). *ROUGE: A Package for Automatic Evaluation of summaries*.

Liu, S., Xu, J., & Wang, H. (2022). Top-down and bottom-up processing in reading comprehension: A review. *Frontiers in Psychology*, *13*, 867531. https://doi.org/10.3389/fpsyg.2022.867531

Louis, A., & Nenkova, A. (2013). Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, *39*(2), 267–300. https://doi.org/10.1162/COLI_a_00123

Louis, A., & Nenkova, A. (2013). Automatically evaluating content selection in summarization without human models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 306–316.

Lu, C. (2011). Automated essay scoring in Chinese: A study on reliability and validity. *Assessing Writing*, *16*(2), 131–146. https://doi.org/10.1016/j.asw.2011.01.001

Martín-Núñez, J. L., Ar, A. Y., Fernández, R. P., Abbas, A., & Radovanović, D. (2023). Does intrinsic motivation mediate perceived artificial intelligence (AI) learning and computational thinking of students during the COVID-19 pandemic? *Computers and Education: Artificial Intelligence*, *4*, 100128. https://doi.org/https://doi.org/10.1016/j.caeai.2023.100128

Meta AI. (2024). *Llama 3.1: Open and efficient foundation language models*. Meta AI. https://huggingface.co/meta-llama

Morris, W., Crossley, S., Holmes, L., Ou, C., Dascalu, M., & McNamara, D. (2024). *Formative Feedback on Student-Authored Summaries in Intelligent Textbooks Using Large Language Models.* International Journal of Artificial Intelligence in Education. Advance online publication. https://doi.org/10.1007/s40593-024-00395-0

Munaye, Y. Y., Admass, W., Belayneh, Y., Molla, A., & Asmare, M. (2025). ChatGPT in Education: A Systematic Review on Opportunities, Challenges, and Future Directions. *Algorithms*, *18*(6), 352. https://doi.org/10.3390/a18060352

Nadea, A. & Jumariati, Jumariati & Nasrullah, Nasrullah. (2021). Bottom-up or Top-down Reading Strategies: Reading Strategies Used by EFL Students. https://doi.org/10.2991/assehr.k.211021.005

Nelson, N. W., & King, J. (2022). Writing summaries: Instructional approaches and challenges. *Journal of Writing Research*, *14*(2), 325–349. https://doi.org/10.17239/jowr-2022.14.02.05

Nieminen, P., & Isohanni, M. (2020). Machine learning approaches for evaluating academic writing: A review. *Journal of Writing Analytics*, *4*, 124–146.

Özdemir, S. (2018). The Effect of Summarization Strategies Teaching on Strategy Usage and Narrative Text

Summarization Successi. *Universal Journal of Educational Research 6*(10): 2199-2209. https://doi.org/10.13189/ujer.2018.061018

P. Lagakis and S. Demetriadis, "Automated essay scoring: A review of the field," 2021 International Conference on Computer, Information and Telecommunication Systems (CITS), Istanbul, Turkey, 2021, pp. 1-6, https://doi.org/10.1109/CITS52676.2021.9618476

Perfetti, C. A. (1985). *Reading ability*. Oxford University Press.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Routledge.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33–58). Lawrence Erlbaum Associates.

Scott Mathews, F. (1985). The structure, function and evolution of cytochromes. *Progress in Biophysics and Molecular Biology*, *45*(1), 1-56. https://doi.org/https://doi.org/10.1016/0079-6107(85)90004-5

Silva, C., & Limongi, R. (2019). Teaching summary writing: A strategy-based approach. *Journal of Applied Linguistics and Language Research*, *6*(1), 214–229.

Somasundaran, S., Burstein, J., & Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. *Proceedings of the First Workshop on Discourse Structure in Machine Translation*, 11–21.

Sung, Y.-T., Chang, K.-E., & Huang, J.-S. (2008). Improving children's reading comprehension and use of strategies through computer-based strategy training. *Computers in Human Behavior*, *24*(4), 1552-1571. https://doi.org/https://doi.org/10.1016/j.chb.2007.05.009

Sung, Y.-T., Chang, K.-E., & Huang, J.-S. (2016). Improving children's summarization ability with computer-assisted learning activities. *Computers & Education*, *92–93*, 316–327. https://doi.org/10.1016/j.compedu.2015.10.010

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*, 5998–6008.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive Computer Support for Writing. *Cognition and Instruction*, *22*(3), 333–362. https://doi.org/10.1207/s1532690xci2203_3

Woods, S., Bixler, R., & Sidner, C. (2017). Computational models of student writing: Current state and future directions. *Journal of Educational Data Mining*, *9*(2), 1–20.

Yaghoobzadeh, Y., & Schütze, H. (2015). Corpus-level fine-grained entity typing using contextual information. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 715–725). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1083

Yang, H., He, Y., Bu, X., Xu, H., & Guo, W. (2023). Automatic Essay Evaluation Technologies in Chinese Writing—A Systematic Literature Review. *Applied Sciences*, *13*(19), 10737. https://doi.org/10.3390/app131910737

From Scarcity to Scalability: Lexicon and Grammar Enhanced Amis to Mandarin Translation with GPT Models

Joseph Lin[1] Kai-ying Lin[2] Hung-Yu Kao[3]

[1]Hsinchu County American School, Hsinchu, Taiwan

[2]Institute of Linguistics, Academia Sinica, Taipei, Taiwan

[3]Dept. of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

lintinghaojoseph@gmail.com limkhaiin@gate.sinica.edu.tw hykao@cs.nthu.edu.tw

Abstract

Machine translation (MT) for low-resource languages remains constrained by extreme data scarcity, making traditional fine-tuning infeasible. This study examines Amis→Mandarin translation as a practical case, leveraging GPT-4o-mini and GPT-5-mini with dictionary integration and grammar-informed prompting. Experiments show that GPT-5-mini, supported by dictionary, achieves usable quality (BLEU-3 ~31, COMET ~78, BLEURT ~71). To address the bottleneck of incomplete dictionaries, we propose *Context-Driven Lexical Augmentation*, which infers Mandarin equivalents for unseen Amis terms from corpus context, raising BLEU-3 to 34 and establishing a stronger basis for semi-automatic corpus generation. These results demonstrate that expanding and refining dictionary provides greater benefits than parameter-intensive fine-tuning in extremely low-resource settings.

We also discuss the performance gap between Amis→Mandarin and Mandarin→Amis translation, attributing it to Amis's morphological complexity and narrower semantic coverage. Overall, our resource-driven strategy offers a scalable pathway toward high-quality MT and corpus expansion, ultimately supporting both linguistic research and language revitalization.

Keywords: Low-Resource Languages, Machine Translation, Prompt Engineering, Data Augmentation

1 Introduction

Taiwan's indigenous languages, widely recognized as the cradle of the Austronesian family (Blust, 2013), are critically endangered. Many are spoken by only a few hundred individuals, and their decline threatens both cultural continuity and linguistic diversity. Developing digital tools to support documentation and revitalization has therefore become an urgent priority. Among such tools, machine translation (MT) holds particular promise, as it can facilitate communication across language communities and accelerate the creation of linguistic resources. However, MT systems typically depend on large-scale parallel corpora, standardized orthographies, and comprehensive dictionary-conditions absent for most indigenous languages. This scarcity firmly categorizes them as low- or no-resource languages, requiring methods that can operate effectively under extreme data constraints.

In this work, we focus on Amis→Mandarin translation as a representative case of low-resource MT. Amis is the most widely spoken indigenous language in Taiwan, yet it remains critically underrepresented in digital resources, like large-scale corpora, standardized lexical tools, or annotated datasets to support NLP tasks like MT. We evaluate two large language models—GPT-4o-mini (small-scale) and GPT-5-mini (mid-tier)—in combination with existing Amis–Mandarin dictionaries (Zheng et al., 2022) and descriptive grammar resources (阿美語語法概論) (Council of Indigenous Peoples, 2017). Our experiments show that grammar-informed prompting benefits smaller models, while GPT-5-mini achieves strong performance with dictionary support alone. Error analysis further reveals that dictionary coverage, rather than syntactic complexity, is the principal bottleneck. To address this, we propose a *Context-Driven Lexical Augmentation* method that infers translations for unseen Amis words, yielding mea-

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 167–175
November 20-22, 2025 ©2025 Association for Computational Linguistics

surable improvements in BLEU and semantic metrics. These findings highlight that systematic enrichment of dictionary is more effective than scaling model size or parallel data alone.

We emphasize the Amis→Mandarin direction for both cultural and practical reasons. From a preservation perspective, initiating data collection from Mandarin and translating into Amis risks cultural bias, as it inevitably introduces Mandarin concepts absent from Amis traditions. Practically, Amis narratives can be translated into Mandarin with sufficient accuracy to bootstrap new corpora while keeping human validation effort low. In contrast, Mandarin encompasses a much broader semantic space, covering domains such as science, technology, and politics, whereas Amis maintains a more compact lexicon rooted in ecology, kinship, and oral tradition (Lewis et al., 2023). As a result, corpora originating in Amis are more coherent and transferable into Mandarin, while the reverse direction often requires paraphrase or approximation that BLEU score penalizes heavily. Taken together, these factors make Amis→Mandarin the most ethical and reliable pathway for semi-automatic corpus expansion.

At the same time, we acknowledge that some prior studies have reported higher BLEU scores for Mandarin→Amis, contrary to our findings. We discuss potential reasons for this discrepancy and its implications for evaluating low-resource MT. Finally, although this study focuses on Amis→Mandarin, we also outline a pathway toward robust bidirectional MT. By combining semi-automatic corpus generation, dictionary augmentation, and semantic-aware evaluation, future work can enable fine-tuning and ultimately achieve high-quality Amis–Mandarin translation in both directions, advancing both NLP research and the revitalization of Taiwan's indigenous languages.

Our contributions are threefold:

- We conduct the first systematic evaluation of Amis→Mandarin translation with GPT-based models, showing that mid-tier LLMs achieve usable quality when supported by dictionary resources.
- We further introduce *Context-Driven Lexical Augmentation*, a lightweight method for expanding dictionary coverage by inferring translations for unseen Amis words, directly improving BLEU and semantic scores.
- We establish Amis→Mandarin as a practical direction for semi-automatic corpus generation and outline a pathway toward high-quality bidirectional MT through future fine-tuning on expanded corpora.

2 Related Work

Low-resource machine translation (MT) faces a fundamental challenge: parallel corpora are too small to support stable fine-tuning of large models. With billions of parameters but only a few thousand sentence pairs, gradient updates are weak, training quickly overfits to idiosyncratic examples, and generalization suffers (Haddow et al., 2022). Empirical studies suggest that with only a few to at most tens of thousands of pairs, full fine-tuning of large models tends to be unstable and prone to overfitting, making parameter-efficient alternatives preferable (Gu et al., 2018). Since the Amis–Mandarin corpus contains only ~5,000 pairs, our setting falls well below this threshold, motivating approaches that leverage external resources such as bilingual dictionaries and grammatical descriptions rather than relying solely on parallel data.

A range of alternatives to full fine-tuning has been explored. Prompt-based methods and in-context learning reduce dependence on large datasets but often deliver inconsistent results. For example, retrieval-augmented prompting with dictionary support reached BLEU ~21 for English–Mambai in one domain but dropped to ~4 in another, revealing limited robustness (Merx et al., 2024). Prompt tuning can exploit structural cues (Schucher et al., 2022), yet its success is highly sensitive to template design and it often struggles to enforce lexical fidelity. Liao et al. (Liao et al., 2024) examined error-feedback prompting for Mandarin→Amis translation, showing that iterative correction brought modest improvements. By contrast, our work centers on Amis→Mandarin, integrating dictionary and grammar resources into prompting and extending coverage through automated lexical augmentation, producing more reliable gains.

In the Formosan and Austronesian context, resources remain sparse but are slowly expanding. Zheng et al. (Zheng et al., 2022) introduced the first Amis–Mandarin parallel corpus and dictionary, demonstrating that dictionary augmentation benefits fine-tuned mBART models. Their experiments reported higher BLEU for Mandarin→Amis (15–19) than for Amis→Mandarin (<7), suggesting directional asymmetry. Yet other research points the other way: Zhang et al. (Zhang et al., 2024) showed that Mandarin→Zhuang achieved much lower BLEU (~16) than Zhuang→Mandarin (~32). Taken together, these studies indicate that directionality may be influenced by morphology, semantic coverage, dictionary completeness, and modeling strategy (fine-tuning vs. prompting).

Lin et al. (Lin et al., 2025) advanced this line of work by releasing FormosanBench, a benchmark spanning Amis, Atayal, and Paiwan across several NLP tasks, including MT. Their evaluation revealed persistent performance gaps relative to high-resource languages, underscoring the importance of approaches tailored to the typological and lexical characteristics of Formosan languages rather than relying exclusively on transfer from unrelated high-resource settings.

Building on this foundation(Lin, 2025), our work proposes a dictionary- and grammar-driven framework for Amis→Mandarin translation with GPT models. Unlike earlier prompting studies that relied on static dictionary , we introduce *Context-Driven Lexical Augmentation*, a proof-of-concept method that infers Mandarin equivalents for unseen Amis words from corpus context. This augmentation improved BLEU-3 from ~31 to 34 and raised semantic scores, surpassing the modest gains reported for prior prompting strategies (Merx et al., 2024; Liao et al., 2024). More broadly, our findings suggest that lexical expansion and semantic-aware evaluation are more scalable and effective than parameter-intensive fine-tuning in extremely low-resource conditions, while also shedding new light on the role of directionality in Amis–Mandarin MT.

3 Translation Framework and Evaluation Metrics

We present an Amis–Mandarin translation framework that integrates mid-tier large language models (LLMs) with lexical and grammatical resources. The system combines dictionary pre-searching and grammar-informed prompting with an iterative auto-prompting procedure, which refines outputs by dynamically adjusting prompts across evaluation rounds. This design offers a practical and scalable strategy for machine translation in low-resource Austronesian languages.

3.1 Models and Data

We evaluate two large language models: GPT-4o-mini, a smaller-scale model, and GPT-5-mini, a mid-tier model. The dataset comprises 5,751 Amis–Mandarin sentence pairs (Zheng et al., 2022), partitioned into 576 for training (used exclusively in the auto-prompt setting), 575 for validation, and 4,600 for testing across all prompt strategies. To maximize evaluation coverage, 80% of the data is allocated to testing, reflecting the fact that prompt engineering does not rely on training sets. In addition, we utilize a bilingual glossary containing 7,927 Amis–Mandarin entries (Zheng et al., 2022), implemented as a Pandas DataFrame to facilitate efficient search and retrieval. Collectively, these resources serve as the most comprehensive Amis–Mandarin parallel dataset currently available.

3.2 Preprocessing

All sentence pairs were standardized before translation. Preprocessing involved removing extra spaces, newline markers, and punctuation. Amis tokens were lowercased while retaining apostrophes, which carry morphological information. Sentences were processed in batches of 20 to balance efficiency with model context length. For each sentence, tokens were normalized and matched against the glossary using RapidFuzz similarity, retrieving up to three candidate translations, or a single match when similarity exceeded 95%.

3.3 Prompting Strategies

We evaluate three prompting strategies, each incorporating glossary-based lexical hints:

1. **Baseline Prompting** (Figure 1(a)): Prompts incorporate detailed formatting instructions and a comprehensive glossary look-up table to guide initial translation efforts. For each word, a fuzzy matching algorithm is employed to retrieve up to three candidate translations, prioritizing those exceeding 80% similarity to ensure high relevance. Sentences are processed efficiently in batches of 20 to optimize computational resources and maintain consistency across translations.

2. **Grammar-Rule Prompting** (Figure 1(b)): Builds on the baseline by appending a curated set of rules from the Amis grammar book (Council of Indigenous Peoples, 2017) on word order, affixation, and case markers, using a similar batching process. Instead of embedding the full 177-page 《秀姑巒阿美語—語法概論》 into each prompt, we use GPT-5 to distill it into a ~3-page "core pack" of high-impact rules (e.g., clause structure, linker *a*, case/voice morphology, negation, relative clauses). An ablation study with 500 randomly selected sentence pairs showed that appending the full dictionary offered no benefit in improving scores. This pack is *frozen* and prepended to Amis→Mandarin prompts. This extract-then-inject approach may mitigate long-context issues like "lost in the middle" (Liu et al., 2024). The compact pack lowers latency/cost, freeing tokens for lexicon snippets and enhancing controllability.

3. **Auto-Prompt Training** (Figure 1(c)): An iterative refinement cycle designed to enhance translation quality through systematic feedback, comprising:
   (a) **Batch Translation**: Process 20-sentence batches using the baseline setup, ensuring consistent input handling and initial translation generation across the corpus.
   (b) **Error Analysis**: Conduct a detailed comparison of translated outputs against reference texts, identifying systematic errors such as lexical mismatches, syntactic deviations, or semantic inaccuracies to pinpoint areas for improvement.
   (c) **Prompt Update**: Revise prompt instructions to address identified issues, incorporating targeted adjustments —e.g., clarifying ambiguous rules or adding contextual cues—based on error patterns observed.
   (d) **Iteration**: Apply the updated prompt to subsequent batches, iteratively refining the process across the training dataset to progressively enhance translation fidelity and coherence.

   Auto-Prompt Training can be conceptualized as a dynamic process wherein the large language model (LLM) implicitly derives grammar rules and linguistic patterns through iterative error analysis and correction. This self-adaptive mechanism leverages accumulated insights to evolve the prompt, with prompts automatically generated by GPT based on the difference between prediction and ground truth reference. The resulting optimized prompt, enriched with corrections and contextual understanding, is then deployed across the test dataset.

3.4 Evaluation Metrics

We assess translation quality using three complementary metrics—BLEU-3, BLEURT, and COMET—each normalized to a 0—100 scale for unified comparison.

BLEU-3, which measures up to 3-gram overlap, is better suited than full BLEU for low-resource MT because shorter n-grams are more reliably captured and impose fewer penalties on valid paraphrases (Liao et al., 2024). The BLEU-3 score is calculated as:

$$\text{BLEU-3} = \text{BP} \cdot \exp\left(\sum_{n=1}^{3} \frac{1}{3} \log p_n\right),$$

where p_n is the precision for each n-gram order and BP is the brevity penalty.

For all BLEU calculations, we apply the `method1` smoothing function from NLTK's `SmoothingFunction`. Under this method, when an n-gram precision would otherwise be zero, it is replaced with a very small constant

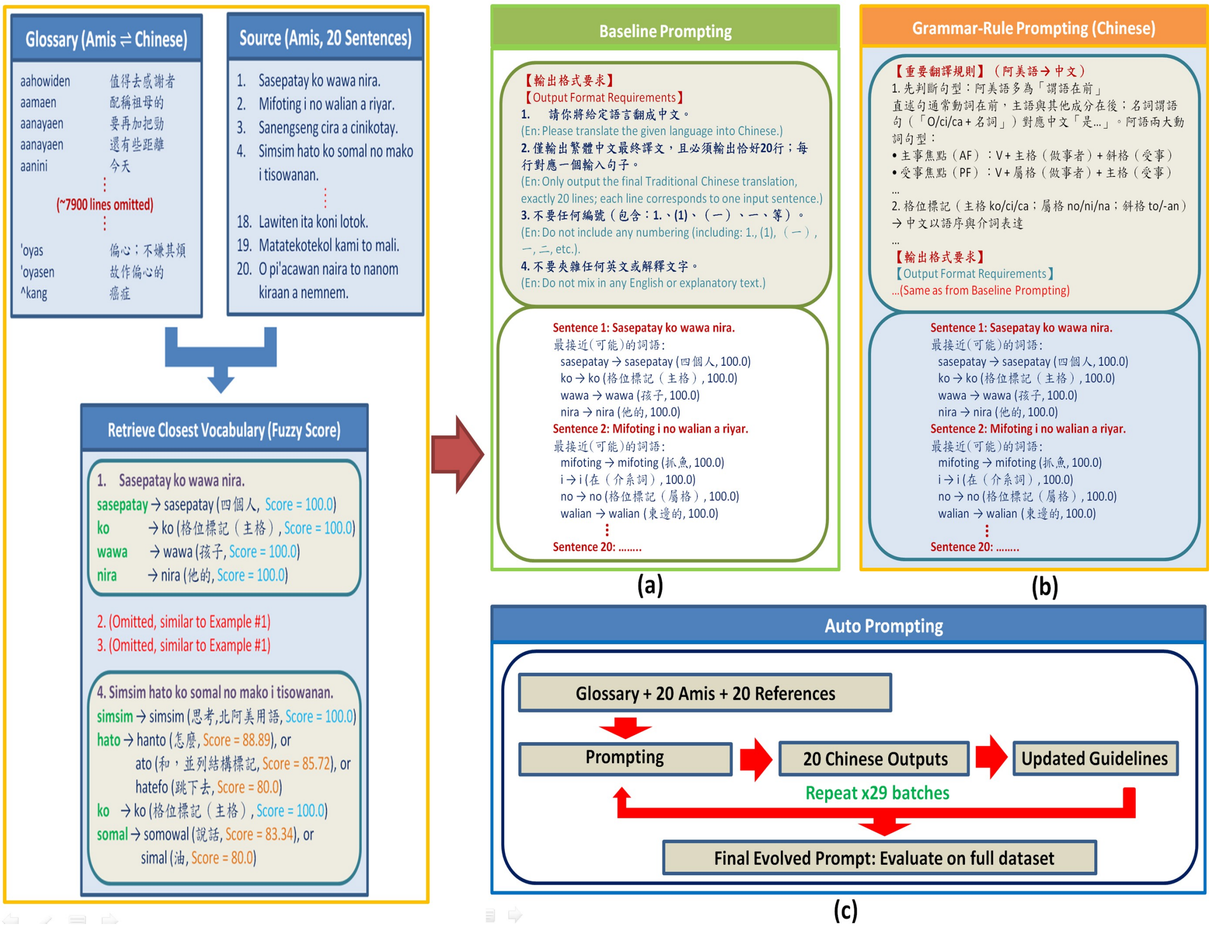


Figure 1: Overview of the prompting framework for Amis→Mandarin MT. Each Amis sentence is first processed through fuzzy matching to retrieve up to three glossary candidates with corresponding Mandarin translations. These lexical hints are then integrated into either (a) **Baseline Prompting**, which applies only formatting requirements, or (b) **Grammar-Rule Prompting**, which supplements the baseline with explicit grammatical rules. (c) **Auto-Generated Prompting** further refines the prompt iteratively by comparing translations against references, analyzing errors, and updating guidelines before final evaluation on the full dataset.

ϵ instead. This avoids BLEU scores collapsing to zero, which is particularly important in our setting where sentences are short and higher-order matches are often sparse. In cases where matches do exist, smoothing has no effect.

To evaluate semantic quality beyond surface overlap, we also include BLEURT and COMET. BLEURT uses pretrained language models fine-tuned on human ratings, while COMET leverages multilingual contextual embeddings; both metrics correlate strongly with human judgments of translation quality. As with BLEU, we normalize BLEURT and COMET to a 0—100 range for consistent comparison.

By combining n-gram precision (BLEU-3 with smoothing) and semantic adequacy (BLEURT, COMET), our evaluation framework balances literal accuracy with meaning preservation—an essential requirement for low-resource MT.

4 Results and Analysis

4.1 Evaluation with GPT Models

Figure 2 presents the performance of GPT-4o-mini and GPT-5-mini under three prompting strategies—Baseline, Grammar-Rule, and Auto-Prompt—plus an additional condition for GPT-5-mini with an augmented dictionary. Evaluation metrics include BLEU1—4, BLEU-3 (our primary n-gram metric for low-resource MT), and the semantic measures COMET and BLEURT, all normalized to a 0—100 scale. Figure 3 plots BLEU-3, COMET, and

BLEURT scores across models and prompting methods for easy comparison.

GPT-4o-mini						
	BLEU1	BLEU2	BLEU3	BLEU4	COMET	BLEURT
Baseline	52	34	20	13	60	61
Grammar-Rule	55	36	23	16	65	66
Auto-Prompt	55	35	22	14	62	64

GPT-5-mini						
	BLEU1	BLEU2	BLEU3	BLEU4	COMET	BLEURT
Baseline	59	43	31	24	78	71
Grammar-Rule	60	43	32	24	78	72
Auto-Prompt	59	43	31	24	77	71
Baseline + Updated Dictionary	62	46	34	26	79	73

Figure 2: Comparison of GPT-4o-mini and GPT-5-mini performance with BLEU1-4, COMET, and BLEURT. Grammar-Rule prompting benefits GPT-4o-mini (BLEU-3: 20→23, COMET: 60→65, BLEURT: 61→66). For GPT-5-mini, dictionary augmentation delivers the largest improvement (BLEU-3: 31→34, COMET: 78→79, BLEURT: 71→73), whereas grammar rules and auto-prompting provide only marginal gains.

For GPT-4o-mini, Grammar-Rule prompting yielded consistent gains over both Baseline and Auto-Prompt: BLEU-3 rose from 20 to 23, COMET from 60 to 65, and BLEURT from 61 to 66. Auto-Prompt achieved a BLEU-3 of 22 but lagged on semantic metrics, indicating that explicit grammatical guidance is particularly valuable for smaller models that struggle with morphosyntactic variation.

GPT-5-mini, by contrast, performed strongly across all conditions, showing that it can already exploit dictionary support without extensive prompting. Baseline results reached BLEU-3 ~31, COMET ~78, and BLEURT ~71. Grammar-Rule and Auto-Prompt strategies offered only marginal gains (BLEU-3 at 32, with semantic scores differing by at most one point), suggesting that larger models are less dependent on handcrafted grammatical cues and generalize robustly from lexical hints alone.

The most notable improvement for GPT-5-mini came from context-driven lexical augmentation (detailed in the next subsection). Incorporating inferred dictionary entries for out-of-vocabulary terms increased BLEU-3 from 31 to 34, COMET from 78 to 79, and BLEURT from 71 to 73. This pattern indicates that dictionary completeness, rather than prompt complexity, is the decisive factor in improving

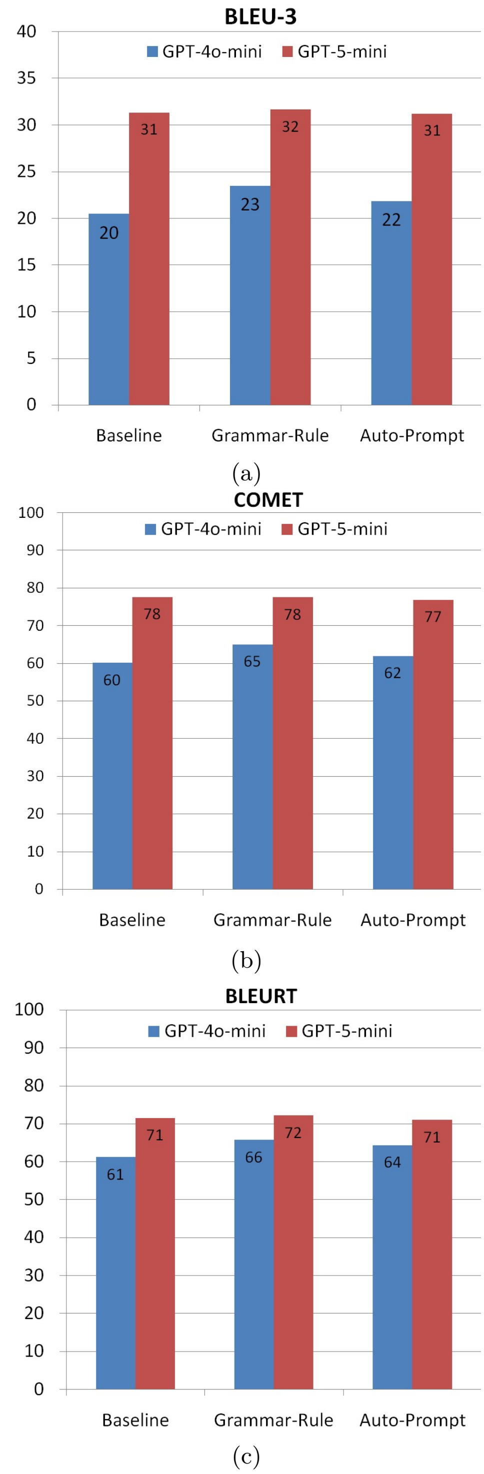


Figure 3: Comparison of GPT-4o-mini and GPT-5-mini performance **before dictionary augmentation**: (a) BLEU-3, (b) COMET, and (c) BLEURT.

translation quality. While additional prompting yields diminishing returns, enriching lexical coverage directly addresses the core bottleneck of low-resource MT.

In sum, GPT-5-mini consistently outperformed GPT-4o-mini on both surface overlap and semantic adequacy. Grammar rules provide clear benefits for smaller models, but for stronger LLMs, the greatest gains derive from expanding dictionary rather than layering increasingly complex prompts.

4.2 Context-Driven Lexical Augmentation

Table 1 demonstrates the effect of out-of-vocabulary (OOV) tokens on translation quality for specific cases. In the baseline system, sentences containing unknown terms such as *fitaol* achieved only BLEU-3 scores of 8一11. To address this, we applied a context-driven lexical augmentation strategy: candidate Mandarin equivalents were inferred from parallel corpus contexts (e.g., *fitaol* → 蛋殼, *pilipayan* → 禮拜天, *atolay* → 地震), and new entries were added to the Amis一Mandarin dictionary, which originally contained ~8000 words, with ~1200 additional OOV words incorporated. For OOV items occurring in multiple sentences, all plausible Mandarin interpretations were appended under the same Amis entry.

We then re-evaluated translation performance with GPT constrained to the augmented dictionary and extracted grammar rules, excluding access to reference translations to preserve evaluation integrity. This approach substantially improved BLEU-3 scores for sentences containing unseen Amis words, with some cases reaching 100 (Table 1). These results highlight the pivotal role of enriched dictionary in improving low-resource MT. While the inferred mappings require further validation by native speakers, the findings underscore the importance of systematic lexical development as a foundation for advancing Amis→Mandarin translation. Future ablation studies could isolate the impact of dictionary size versus prompt complexity to further refine these gains.

Amis (Reference below)	**Baseline (BLEU-3)**	**Updated Dictionary (BLEU-3)**
Mifitelak to fitaol ko ciwciw。 *Reference:* 小雞破殼而出。	小雞把蛋弄破了。 (11)	小雞破殼而出。 (100)
O pilipayan i nacila。 *Reference:* 昨天是禮拜天。	那是昨天發生的事。 (9)	昨天是禮拜天。 (100)
Mangernnger no atolay ko loma'。 *Reference:* 房屋被地震震動。	我家的南邊被震動了。 (8)	房屋被地震震動。 (100)

Table 1: Impact of context-driven lexical augmentation on Amis–Mandarin translation. Augmentation resolves OOV terms (e.g., *fitaol* → 蛋殼, *pilipayan* → 禮拜天, *atolay* → 地震), improving BLEU-3 from 8一11 to up to 100 for specific cases.

4.3 Limitations of BLEU for Semantic Evaluation

Table 2 illustrates cases where BLEU underestimates translation quality. GPT-5-mini Auto-Prompt outputs for some sentences receive very low BLEU scores (4-8) despite being semantically accurate, as reflected by much higher COMET scores (72一90). For instance, the Amis sentence *"Narikoran no faliyos matomes ko sota' i lalan."* is translated as " 颱風過後，路上滿是泥巴。" (BLEU-3 = 8, COMET = 90). While lexically divergent from the reference, the meaning is preserved.

Similar discrepancies appear in other examples, where paraphrasing reduces BLEU but COMET captures semantic fidelity. These results highlight BLEU's limitations in low-resource MT, particularly for languages where flexible phrasing is common. Semantic metrics such as COMET provide better alignment with human judgment and should complement BLEU in evaluation frameworks for endangered and low-resource languages. Future evaluations might explore large language models for direct scoring to further reduce bias.

Amis	**Reference / (GPT-5-mini)**	**COMET (BLEU-3)**
Narikoran no faliyos matomes ko sota’i lalan.	颱風之後馬路填滿了污泥。 （颱風過後，路上滿是泥巴。）	90 (8)
Aka pahacikay a miparakat to tosiya.	不可開快車。 （不要把車開得太快。）	88 (5)
Do^do han ko rakat ako!	請跟從我的腳步！ （跟著我走！）	84 (4)
Ma’adangen kako to ngiha’no dadacdac.	我覺得蟬叫聲很吵。 （我被蟬的聲音吵到。）	72 (5)

Table 2: Examples where BLEU-3 penalizes paraphrasing despite high semantic fidelity, as reflected by COMET.

5 Challenges and Pathways for Amis–Mandarin MT

Prior work has reported that Mandarin→Amis translation can achieve higher BLEU than the reverse. In contrast, our experiments consistently find the opposite: Amis→Mandarin yields stronger performance. We attribute this to the fact that Amis is morphologically rich, with many surface forms for the same concept, which leads to frequent mismatches under BLEU. For example, a single Mandarin word may correspond to several Amis forms depending on context, and without explicit disambiguation, models often choose the wrong variant, resulting in lower BLEU score.

These challenges highlight why dictionary expansion with contextual metadata are crucial for future progress. Our system already achieves sufficient accuracy in the Amis→Mandarin direction to enable large-scale semi-automatic corpus generation, easing the burden on human validators and accelerating resource development. By complementing this with automatic dictionary augmentation, we can steadily improve lexical coverage and translation fidelity.

Looking ahead, we see a clear pathway: use automatic Amis→Mandarin translation to bootstrap corpora from elder narratives, refine outputs through lightweight human feedback, and progressively enrich the dictionary with contextual information. In the long term, integrating direct speech-to-text translation will further reduce barriers to language documentation and revitalization, while offering a generalizable framework for other low-resource, morphologically complex languages.

6 Conclusion

This study examined Amis→Mandarin translation as a practical case of low-resource MT, focusing on strategies that enable scalable corpus expansion despite the limited parallel data available. Our experiments show that mid-tier LLMs, particularly GPT-5-mini, can achieve usable quality when paired with dictionary support (BLEU-3 ~31, COMET ~78, BLEURT ~71). The proposed framework is applicable to any large language model comparable to or exceeding the capabilities of GPT-5-mini. While grammar-informed prompting benefits smaller models, dictionary coverage emerged as the decisive factor.

We further demonstrated that augmenting the glossary with context-inferred entries improves translation quality and establishes a threshold where large-scale semi-automatic data generation becomes feasible. This approach allows Amis narratives to be translated into Mandarin with sufficient accuracy for bootstrapping new corpora, reducing human effort to lightweight validation.

In summary, our contribution lies in reframing low-resource MT for endangered languages: progress is driven less by parameter-intensive fine-tuning and more by systematic lexical expansion, context-sensitive dictionary design, and semantic-aware evaluation. Crucially, the resulting expanded corpora will make future fine-tuning feasible, enabling higher-quality Amis–Mandarin bidirectional MT and providing a sustainable foundation for language preservation.

Acknowledgements

This research is supported by the Institute of Linguistics, Academia Sinica, Taiwan, and the AI Advanced Talent Cultivation Program of Taiwan's Ministry of Education.

References

Robert Blust. 2013. *The Austronesian Languages*. Asia-Pacific Linguistics.

Council of Indigenous Peoples. 2017. *臺灣南島語言叢書_1. 阿美語語法概論*. Council of Indigenous Peoples, Taipei, Taiwan.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of NAACL-HLT*, pages 344–354.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Molly Lewis, Aoife Cahill, Nitin Madnani, and James Evans. 2023. Local similarity and global variability characterize the semantic space of human languages. *Proceedings of the National Academy of Sciences*, 120(51):e2300986120.

You Cheng Liao, Chen-Jui Yu, Chi-Yi Lin, He-Feng Yun, Yen-Hsiang Wang, Hsiao-Min Li, and

Yao-Chung Fan. 2024. Learning-from-mistakes prompting for indigenous language translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 146–158, Bangkok, Thailand. Association for Computational Linguistics.

Joseph Lin. 2025. Tackling data scarcity: A practical framework for amis-to-mandarin machine translation. The Fourth Taiwan High School Linguistics Science Fair, National Taiwan Normal University, Taipei, Taiwan.

K. K. Lin, H. Chen, and H. Zhang. 2025. Formosanbench: Benchmarking low-resource austronesian languages in the era of large language models. *arXiv preprint arXiv:2506.21563*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Nathan Schucher, Siva Reddy, and Harm de Vries. 2022. The power of prompt tuning for low-resource semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 148–156, Dublin, Ireland. Association for Computational Linguistics.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800, Bangkok, Thailand. Association for Computational Linguistics.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2022. A parallel corpus and dictionary for Amis-Mandarin translation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 79–84, Taipei, Taiwan. Association for Computational Linguistics.

CLiFT-ASR: A Cross-Lingual Fine-Tuning Framework for Low-Resource Taiwanese Hokkien Speech Recognition

Hung-Yang Sung[1], Chien-Chun Wang[1], Kuan-Tang Huang[1], Tien-Hong Lo[1], Yu-Sheng Tsao[2], Yung-Chang Hsu[2], and Berlin Chen[1]
[1]National Taiwan Normal University, Taiwan
[2]EZAI, Taiwan
{redsheep, 61247033s, 61347002s, teinhonglo, berlin}@ntnu.edu.tw,
{sam, mic}@ez-ai.com.tw

Abstract

Automatic speech recognition (ASR) for low-resource languages such as Taiwanese Hokkien is difficult due to the scarcity of annotated data. However, direct fine-tuning on Han-character transcriptions often fails to capture detailed phonetic and tonal cues, while training only on romanization lacks lexical and syntactic coverage. In addition, prior studies have rarely explored staged strategies that integrate both annotation types. To address this gap, we present CLiFT-ASR, a cross-lingual fine-tuning framework that builds on Mandarin HuBERT models and progressively adapts them to Taiwanese Hokkien. The framework employs a two-stage process in which it first learns acoustic and tonal representations from phonetic Tai-lo annotations and then captures vocabulary and syntax from Han-character transcriptions. This progressive adaptation enables effective alignment between speech sounds and orthographic structures. Experiments on the TAT-MOE corpus demonstrate that CLiFT-ASR achieves a 24.88% relative reduction in character error rate (CER) compared with strong baselines. The results indicate that CLiFT-ASR provides an effective and parameter-efficient solution for Taiwanese Hokkien ASR and that it has potential to benefit other low-resource language scenarios.

Keywords: Automatic speech recognition, low-resource language, Taiwanese Hokkien, cross-lingual transfer, two-stage fine-tuning

1 Introduction

Taiwanese Hokkien is an important dialect in Taiwan with rich cultural and historical significance. However, as Mandarin Chinese dominates education and daily life, the use of Taiwanese Hokkien has been declining, especially among younger generations. A 2020 survey[1] reports that only 7.4% of children regularly use Taiwanese Hokkien. Despite the existence of several speech corpora (Liao et al., 2022; Chou et al., 2023; Lin et al., 2024), the overall amount of annotated data is limited compared to high-resource languages such as Mandarin and English (Zhang et al., 2022; Wang et al., 2021). This data scarcity poses a significant challenge for developing robust Speech Translation (Chen et al., 2023) and automatic speech recognition (ASR) systems.

Existing Taiwanese Hokkien ASR systems face additional challenges due to inconsistent transcription standards. Some systems employ Tai-lo romanization (Chou et al., 2023; Chao et al., 2021), which combines phonetic scripts with tonal markings, making it less intuitive and harder for general users to accept (Khoo, 2019). Other approaches annotate speech with Mandarin characters, but the mapping between Taiwanese Hokkien vocabulary and Mandarin text is often one-to-many or partially aligned, leading to longer and less accurate output sequences (Lin et al., 2024). Using Taiwanese Hokkien Han characters provides a practical alternative that balances readability and phonological detail, improving recognition usability.

To overcome these challenges, we introduce **CLiFT-ASR**[2], a **C**ross-**Li**ngual **F**ine-**T**uning framework for low-resource **A**utomatic **S**peech **R**ecognition that leverages Mandarin HuBERT backbone models and progressively adapts them to Taiwanese Hokkien. The framework follows a two-stage fine-tuning

[1]https://www.stat.gov.tw/News_Content.aspx?Create=1&n=2755&state=1327FD6AD8DCDA52&s=230300&ccms_cs=1&sms=11065/

[2]https://github.com/redsheep913/CLiFT-ASR/

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 176–183
November 20-22, 2025

strategy where it first acquires acoustic-level knowledge from phonetic Tai-lo annotations and then learns language-level structures such as vocabulary and syntax using Taiwanese Hokkien Han characters. Comprehensive experiments on the TAT-MOE corpus demonstrate that CLiFT-ASR achieves a 24.88% relative reduction in character error rate (CER). The framework offers an effective solution for Taiwanese Hokkien ASR and provides guidance for developing ASR systems for other low-resource languages.

2 Background

2.1 Linguistic Characteristics of Taiwanese Hokkien

Taiwanese Hokkien has a seven-tone system and exhibits tone sandhi, which creates tonal variations that differ from Mandarin (Cheng, 1968). These tonal patterns make automatic speech recognition challenging, as accurate recognition requires modeling both static tones and context-dependent tone changes. Despite these differences, Taiwanese and Mandarin share similar morphological and syntactic structures (Sun, 2006), which allows knowledge transfer from Mandarin-pretrained models. Previous studies show that Mandarin-pretrained ASR models outperform English-pretrained models when recognizing romanized Taiwanese (Tai-lo), indicating that cross-lingual transfer can be effective for end-to-end ASR targeting Taiwanese Han characters (Chou et al., 2023). These observations motivate the design of CLiFT-ASR, which leverages cross-lingual knowledge and adapts it progressively to Taiwanese Hokkien. Note that this work does not focus on modeling tone sandhi phenomena, which is left for future research.

2.2 Orthographic Systems and Their Role in ASR

Taiwanese Hokkien uses two main orthographies: romanization and Han characters. Romanization systems such as Peh-ōe-jī (POJ) and Tai-lo provide systematic phonetic representations (Khoo, 2019). The Ministry of Education has published a recommended set of roughly 700 Han characters for writing Taiwanese, which can be combined with romanization in a mixed-script form known as hàn-lô [3]. For ASR, using Han characters or hàn-lô offers a practical balance between phonetic detail and readability and informs the two-stage fine-tuning strategy employed in CLiFT-ASR.

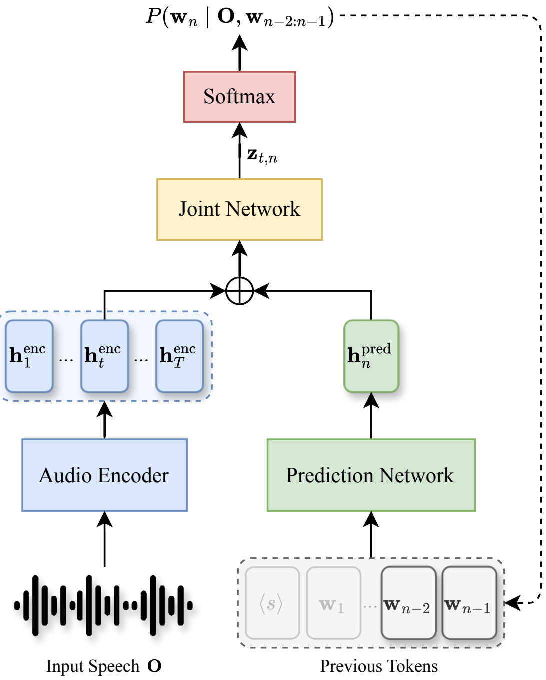


Figure 1: Overview of the proposed CLiFT-ASR. The $\bigoplus$ operator denotes element-wise tensor addition. The dashed arrow indicates that during inference, ground-truth labels are not available, so the model outputs are fed back autoregressively into the prediction network.

3 Proposed Method

3.1 Model Architecture

The proposed CLiFT-ASR framework builds upon the RNN-Transducer (RNN-T) (Graves, 2012) to align variable-length acoustic sequences with token sequences, as illustrated in Figure 1. Given an input audio signal $\mathbf{O}$ and a sequence of target tokens $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)$, where N denotes the number of output tokens, CLiFT-ASR estimates a probability distribution over possible tokens at each alignment step. The model consists of three components: an audio encoder, a prediction network, and a joint network. The audio encoder processes T acoustic feature frame vectors $(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$ extracted from $\mathbf{O}$ and maps them to a sequence of high-level representations that capture phonetic, tonal, and

[3] https://language.moe.gov.tw/.../D005.pdf

other essential speech information:

$$\mathbf{H}^{\text{enc}} = \text{AudioEncoder}(\mathbf{O}), \tag{1}$$

where $\mathbf{H}^{\text{enc}}$ represents the sequence of encoder hidden states $(\mathbf{h}_1^{\text{enc}}, \ldots, \mathbf{h}_T^{\text{enc}})$. This representation integrates both local and global acoustic patterns, which is crucial for accurately modeling the tonal variations in Taiwanese Hokkien. The prediction network generates a context representation autoregressively, conditioning on the previous two target tokens to form a trigram-style context that models short-term sequential dependencies in the output space:

$$\mathbf{h}_n^{\text{pred}} = \text{PredictionNetwork}(\mathbf{w}_{n-2}, \mathbf{w}_{n-1}). \tag{2}$$

The joint network combines the encoder output at time step t, $\mathbf{h}_t^{\text{enc}}$, which corresponds to the t-th element of $\mathbf{H}^{\text{enc}}$, with the prediction network state to form a joint representation. The conditional distribution over the next token is then obtained by applying a softmax:

$$\mathbf{z}_{t,n} = \text{JointNetwork}(\mathbf{h}_t^{\text{enc}} + \mathbf{h}_n^{\text{pred}}), \tag{3}$$

$$P(\mathbf{w}_n \mid \mathbf{O}, \mathbf{w}_{n-2:n-1}) = \text{Softmax}(\mathbf{z}_{t,n}). \tag{4}$$

This architecture allows CLiFT-ASR to jointly leverage acoustic and linguistic context at each step, which is essential for capturing tonal and phonological patterns in Taiwanese Hokkien.

3.2 Training Strategy

To handle limited annotated Taiwanese Hokkien data, CLiFT-ASR adopts a two-stage fine-tuning framework based on a pre-trained Mandarin HuBERT encoder. In the first stage, the model learns acoustic-level representations from phonetic Tai-lo annotations. Given a training set $\{(\mathbf{O}^{(i)}, \mathbf{W}_{\text{Tai-lo}}^{(i)})\}_{i=1}^{U_{\text{Tai-lo}}}$, the model parameters θ are updated to minimize the negative log-likelihood:

$$\theta' = \arg\min_{\theta} \sum_{i=1}^{U_{\text{Tai-lo}}} -\log P\big(\mathbf{W}_{\text{Tai-lo}}^{(i)} \mid \mathbf{O}^{(i)}; \theta\big). \tag{5}$$

This stage captures fine-grained acoustic and phonetic details, providing a solid foundation for language-level learning. In the second stage, the model is fine-tuned on Taiwanese Hokkien Han character annotations. Starting from the network parameter θ', it is trained on $\{(\mathbf{O}^{(j)}, \mathbf{W}_{\text{Han}}^{(j)})\}_{j=1}^{U_{\text{Han}}}$ to learn vocabulary, syntax, and higher-level linguistic structures:

$$\theta^* = \arg\min_{\theta} \sum_{j=1}^{U_{\text{Han}}} -\log P\big(\mathbf{W}_{\text{Han}}^{(j)} \mid \mathbf{O}^{(j)}; \theta\big). \tag{6}$$

By progressively adapting from phonetic to linguistic representations, CLiFT-ASR effectively leverages cross-lingual knowledge and maximizes the use of limited annotated data, resulting in more accurate and robust Taiwanese Hokkien ASR.

Split	**Spk.**	**Utt.**	**Hr.**
Training	328	86,072	153.33
Development	58	16,357	28.60
Test	54	15,962	26.28
Total	440	118,391	208.21

Table 1: Statistics of the TAT-MOE dataset across training, development, and test splits, including the number of speakers (Spk.), utterances (Utt.), and total duration in hours (Hr.).

4 Experimental Setup

4.1 Dataset

All experiments were conducted on the TAT-MOE subset of the TAT corpus (Liao et al., 2022), a large-scale Taiwanese Hokkien speech resource covering diverse regions of Taiwan. The corpus captures variation in speaker accents and pronunciation, providing a suitable testbed for robust ASR development. Audio recordings were sampled at 16 kHz with 16-bit PCM encoding to ensure consistent acoustic quality. Transcriptions were provided in Hàn-Lô-Tâi-bûn, a mixed orthography combining Han characters and romanized phonetics. Alternative annotations, including Peh-ōe-jī, Tai-lo, and tone-marked Tai-lo, were also available to support different modeling strategies. Table 1 summarizes the number of speakers, utterances, and total duration for the training, development, and test sets. To further evaluate model performance, we included a cleaner test set drawn from the pilot test of the Formosa Speech Recognition Challenge 2020 (FSR-2020) (Liao et al., 2020), referred to as the clean test. The TAT-MOE corpus therefore provides high-quality acoustic data and multiple orthographic representations, making

Model	Parameters (M)	Development		Test		Clean Test	
		CER	Rel.	CER	Rel.	CER	Rel.
Zipformer	65	48.57	-	45.82	-	15.69	-
FSR-2020 Best	-	-	-	-	-	15.62	0.07
Whisper-base	74	27.36	21.21	24.02	21.80	10.05	5.64
HuBERT-base	96	26.16	22.41	24.49	21.33	12.97	2.72
HuBERT-base-cmn	96	24.06	24.51	22.41	23.41	9.08	6.61
CLiFT-ASR	**96**	**22.37**	**26.20**	**20.94**	**24.88**	**8.06**	**7.63**
Whisper-small	244	22.47	26.10	18.68	27.14	7.66	8.03

Table 2: CERs (%) and relative reductions (Rel., %) for Taiwanese Hokkien ASR using various audio encoder initialization strategies.CLiFT-ASR applies a two-stage fine-tuning strategy with the HuBERT-base-cmn encoder. FSR-2020 Best refers to the top-performing model from FSR-2020.

it a valuable benchmark for low-resource Taiwanese Hokkien ASR.

4.2 Data preprocessing

The transcripts in the TAT-MOE dataset were written in Hàn-Lô-Tâi-bûn, a mixed system of Han characters and romanized phonetics. To unify the representation, we first constructed a mapping table using additional corpora to convert romanized segments into the corresponding Han characters. Arabic numerals were also converted into Chinese numerals, and variant or synonymous characters were normalized to a single standardized form. These preprocessing steps reduce inconsistencies and lexical variation in the annotations, thereby improving the stability of training and the accuracy of recognition.

4.3 Model Configuration

All training procedures followed Icefall's official recipes and default settings[4]. To establish a fair baseline and assess the benefit of cross-lingual transfer, we considered two encoder configurations: the baseline Zipformer model and a HuBERT-based Transformer initialized with Mandarin pretrained weights provided by the toolkit. The prediction network adopted Icefall's stateless design for efficient sequence modeling, and the joint network followed the standard implementation for integrating audio encoder and prediction network features into output distributions (Yao et al., 2024; Hsu et al., 2021; Ghodsi et al., 2020). For tokenization, we employed Icefall's byte-level BPE model, which has proven effective for handling large CJK vocabularies in bilingual and multilingual ASR tasks. This configuration enables a direct comparison between a strong baseline and our cross-lingual strategy, ensuring that performance gains are consistent and interpretable.

4.4 Training Details

Speech data were prepared using the Lhotse toolkit (Żelasko et al., 2021). For feature extraction, the Zipformer baseline model used 80-dimensional filter bank (FBank) features, while the HuBERT-based model was fine-tuned directly from raw waveform inputs. In CLiFT-ASR, the first stage was trained for 20 epochs and the second stage for 40 epochs. For comparison, a direct fine-tuning approach without staging was trained for 60 epochs. All models were trained with gradient accumulation over 4 steps to stabilize optimization. To balance computational efficiency and contextual coverage, the maximum audio duration per training sample was limited to 120 seconds. The learning rate was initialized at 0.0005 and scheduled over 40 epochs for smooth convergence. Model embeddings were set to 256 dimensions, and training was initialized from pretrained checkpoints. Optimization was performed with the ScaledAdam optimizer, which applied adaptive learning rates and gradient clipping at 2.0 for stability. A custom learning rate scheduler, Eden, was employed to dynamically adjust the learning rate across both batch and epoch progression (Yao et al., 2024).

[4] https://github.com/k2-fsa/icefall/

Fine-tuning Strategy	Frozen	Development	Test	Clean Test
Direct	None	24.06	22.41	9.08
Two-stage	Audio Encoder	36.82	35.72	26.84
	Prediction Network	25.23	23.91	11.84
	Joint Network	29.58	28.59	18.46
	None	**22.37**	**20.94**	**8.60**

Table 3: CERs (%) on development, test, and clean test sets for different training strategies and parameter freezing configurations. The table compares direct fine-tuning with the proposed two-stage strategy, evaluating the impact of freezing specific components (audio encoder, prediction network, joint network) during the first stage.

4.5 Evaluation Metric

Character error rate (CER) was employed as the primary evaluation metric. CER quantifies the discrepancy between the predicted output and the reference transcription by counting the number of substitutions, deletions, and insertions. It is computed as the ratio of total character errors to the number of characters in the reference:

$$\mathrm{CER} = \frac{S + D + I}{C}, \tag{7}$$

where S, D, and I represent the numbers of substitutions, deletions, and insertions, respectively, and C denotes the total number of characters in the reference. As a character-level measure, CER provides a precise and widely accepted evaluation of recognition accuracy for speech recognition tasks, with lower values indicating better performance. For Taiwanese Hokkien, where annotations include a mix of Han characters and romanized phonetics, CER is particularly suitable because it captures errors across both orthographic forms and effectively reflects the ability of the model to handle tonal and phonological variations.

5 Results and Discussion

5.1 Effects of Language Initialization

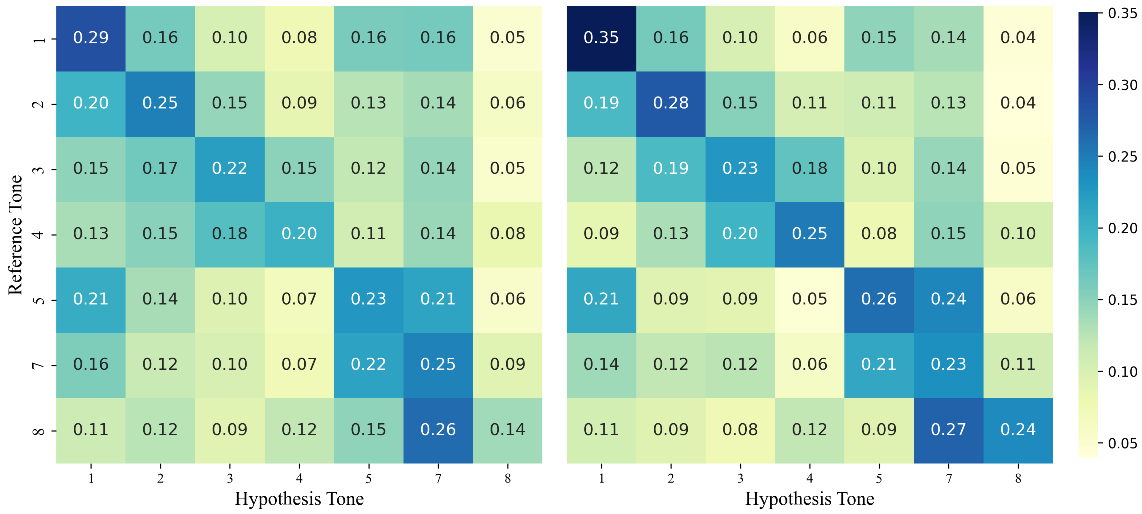
Table 2 summarizes the impact of different encoder initialization strategies on Taiwanese Hokkien ASR performance. The comparison includes Zipformer without pretraining, HuBERT-base pretrained on English, Whisper models with multilingual pretraining, and HuBERT-base-cmn pretrained on Mandarin. CLiFT-ASR, built on the Mandarin-pretrained HuBERT-base-cmn encoder and the proposed two-stage fine-tuning strategy, achieves the strongest overall performance.

Models without language-specific pretraining, such as Zipformer, exhibit the lowest performance, highlighting the difficulty of learning effective acoustic representations from limited Taiwanese data alone. Whisper-base, benefiting from large-scale multilingual pretraining, shows significant improvement and robust generalization across languages. English-pretrained HuBERT-base offers moderate gains, indicating that cross-lingual transfer helps but is constrained by the phonological mismatch between English and Taiwanese Hokkien. Compared with these strong baselines, CLiFT-ASR consistently reduces CER across all evaluation sets, achieving up to 26.2% relative improvement on the development set and 24.88% on the test set. While Whisper-small slightly outperforms CLiFT-ASR on certain splits, it contains more than twice the number of parameters. CLiFT-ASR therefore offers a parameter-efficient solution with substantial gains over competitive baselines, demonstrating the effectiveness of cross-lingual initialization combined with progressive two-stage fine-tuning.

5.2 Analysis of Fine-tuning Strategies

Table 3 presents the effects of different fine-tuning strategies and parameter freezing configurations on CLiFT-ASR performance. Compared with direct end-to-end fine-tuning, the proposed two-stage strategy, which first adapts the model on phonetic (romanized) transcriptions and then refines it with Han character targets, consistently improves recognition accuracy across all evaluation sets.

Analyzing parameter freezing during the



(a) Baseline (b) CLiFT-ASR

Figure 2: Row-normalized substitution confusion matrices for Taiwanese Hokkien tone prediction, comparing the baseline and proposed models. Tone labels are derived from Taibun surface forms without applying sandhi rules.

first stage highlights the contribution of each module. Freezing the audio encoder or joint network restricts the ability of the model to adapt to target phonetics and orthography, leading to notable performance degradation, whereas freezing the prediction network has a milder effect. The lowest CER is achieved when all components are trainable, indicating that full model adaptation within the two-stage fine-tuning strategy enables effective integration of acoustic and linguistic knowledge. These results demonstrate that CLiFT-ASR with a carefully designed multi-stage fine-tuning strategy outperforms direct adaptation and provides a robust solution for low-resource mixed-orthography ASR scenarios.

5.3 Investigation of Tone Confusions

Figure 2 depicts the substitution confusion matrix for tone prediction in Taiwanese Hokkien. The diagonal dominance indicates that most tones are correctly classified, yet tones 5, 7, and 8 exhibit frequent mutual misclassifications. These errors are likely attributed to tone sandhi phenomena, overlapping pitch contours, and speaker-dependent prosodic variations, which complicate accurate tone modeling in ASR. To conduct this analysis, we employed the Taibun tool[5] to convert Taiwanese Han character outputs into Romanized forms with numerical tone labels. By aligning reference and predicted tone sequences, we constructed row-normalized substitution matrices to quantify tone-level confusions.

In the Zipformer baseline, tones 1, 5, and 7 emerge as the most error-prone categories. Tone 1 is correctly recognized only 29% of the time, with 16% of its instances misclassified as tone 2. Tone 5 is frequently misclassified as tone 1 (21%), while tones 7 and 8 show substantial cross-confusions, indicating the limited ability of the baseline model to discriminate between acoustically similar tones. In contrast, the proposed CLiFT-ASR system demonstrates clear improvements across most tonal categories. Tone 1 accuracy increases from 29% to 35%, while tone 4 recognition improves from 20% to 25%. The overall misclassification rate decreases, particularly for tones 5 and 7, where cross-tone errors are substantially reduced. These results highlight the enhanced discriminative capability of the proposed framework. In summary, the tone confusion analysis confirms that CLiFT-ASR effectively reduces inter-tone errors, especially among acoustically similar tone pairs. This improvement can be attributed to the proposed feature design and training strategy, which together provide more robust tonal modeling for Taiwanese Hokkien ASR.

[5] https://github.com/andreihar/taibun/

6 Conclusion

This study presents CLiFT-ASR, a cross-lingual fine-tuning framework designed for low-resource Taiwanese Hokkien ASR. By initializing the audio encoder with Mandarin speech representations and applying an effective two-stage fine-tuning strategy, CLiFT-ASR achieves the best overall performance. The first stage leverages Taiwanese romanization to capture detailed phonetic information, and the second stage adapts to Han character transcriptions to integrate orthographic and syntactic knowledge. This progressive strategy highlights the advantage of aligning acoustic and linguistic representations in stages rather than directly training with limited annotated data. An analysis of tone recognition shows that while general tone recognition is accurate, tones 5, 7, and 8 remain difficult due to tone sandhi, overlapping acoustic patterns, and speaker-specific prosodic variation, all of which complicate precise tone modeling.

7 Future Work

Several directions can be explored to extend the proposed CLiFT-ASR. One promising avenue is targeted data augmentation that balances underrepresented tones. Another is explicit modeling of tone sandhi, which may further reduce tonal confusion. The integration of larger and more diverse pretraining corpora is expected to improve robustness, particularly for conversational speech. Future research may also apply advanced sequence modeling or structured prediction techniques to capture tonal dependencies more effectively. Finally, evaluating multilingual models such as Whisper could provide additional gains through large-scale pre-training and enhanced contextual modeling.

8 Limitations

Although CLiFT-ASR achieves competitive improvements, the current design relies on a stateless RNN-Transducer framework. The stateless prediction network constrains the ability to model long-range dependencies, which may reduce accuracy in recognizing tonal patterns and complex tone sandhi. Compared with recent large-scale pretrained models, the architecture also has limited capacity to exploit fully contextualized acoustic representations. These limitations suggest that adopting more expressive architectures with stronger context modeling could further advance Taiwanese Hokkien ASR.

References

Fu-An Chao, Tien-Hong Lo, Shi-Yan Weng, Shih-Hsuan Chiu, Yao-Ting Sung, and Berlin Chen. 2021. The NTNU Taiwanese ASR system for formosa speech recognition challenge 2020. In *Proc. IJCLCLP*.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, Hirofumi Inaguma, Sravya Popuri, Changhan Wang, Juan Pino, Wei-Ning Hsu, and Ann Lee. 2023. Speech-to-speech translation for a real-world unwritten language. In *Findings of ACL*.

Robert L. Cheng. 1968. Tone sandhi in Taiwanese. *Linguistics*, 6(41):19–42.

Yi-Hui Chou, Kalvin Chang, Meng-Ju Wu, Winston Ou, Alice Wen-Hsin Bi, Carol Yang, Bryan Y. Chen, Rong-Wei Pai, Po-Yen Yeh, Jo-Peng Chiang, Iu-Tshiann Phoann, Winnie Chang, Chenxuan Cui, Noel Chen, and Jiatong Shi. 2023. Evaluating self-supervised speech models on a Taiwanese Hokkien corpus. In *Proc. IEEE ASRU*.

Mohammadreza Ghodsi, Xiaofeng Liu, James Apfel, Rodrigo Cabrera, and Eugene Weinstein. 2020. RNN-transducer with stateless prediction network. In *Proc. ICASSP*.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. In *Proc. ICML*.

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 Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Hui-lu Khoo. 2019. The dynamics of Southern Min in Taiwan: From Southern Min dialects to “Taigi” . In Chris Shei, editor, *The Routledge Handbook of Chinese Discourse Analysis*, pages 596–610. Routledge.

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 *Proc. O-COCOSDA*.

Yuan-Fu Liao, Jane S. Tsay, Peter Kang, Hui-Lu Khoo, Le-Kun Tan, Li-Chen Chang, Un-Gian Iunn, Huang-Lan Su, Tsun-Guan Thiann, Hak-Khiam Tiun, and Su-Lian Liao. 2022. Taiwanese across Taiwan corpus and its applications. In *Proc. O-COCOSDA*.

Jiayan Lin, Shenghui Lu, Hukai Huang, Wenhao Guan, Binbin Xu, Hui Bu, Qingyang Hong, and Lin Li. 2024. MinSpeech: A corpus of Southern Min dialect for automatic speech recognition. In *Proc. Interspeech*.

Chaofen Sun. 2006. *Chinese: A linguistic introduction*. Cambridge University Press.

Xiong Wang, Zhuoyuan Yao, Xian Shi, and Lei Xie. 2021. Cascade RNN-transducer: Syllable based streaming on-device Mandarin speech recognition with a syllable-to-character converter. In *Proc. IEEE SLT*.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2024. Zipformer: A faster and better encoder for automatic speech recognition. In *Proc. ICLR*.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. WenetSpeech: A 10000+ hours multi-domain Mandarin corpus for speech recognition. In *Proc. ICASSP*.

Piotr Żelasko, Daniel Povey, Jan Yenda Trmal, and Sanjeev Khudanpur. 2021. Lhotse: A speech data representation library for the modern deep learning ecosystem. In *Proc. NeurIPS*.

以大語言模型進行兒童敘事句法能力檢測與分析
MINAS: Mandarin Intelligent Narrative Assessment of Syntax for Children

王睿孺 Ruei-Ru Wang[1,*] , 李亞欣 Ya-Sin Li[1,*] , 尹懌碩 Yi-Shuo Yin[1] ,
陳韜宇 Tao-Yu Chen[1] , 張顯達 Hint-Tat Cheung[2] , 陳鯨太 Ching-Tai Chen[3,†]

ctchen@utaipei.edu.tw
*These authors contributed equally to this work.
†corresponding author
1 Department of Bioinformatics and Medical Engineering, Asia University
2 Department of Audiology and Speech-Language Pathology, Asia University
3 Department of Computer Science, University of Taipei

摘要

兒童敘事能力是語言發展的重要指標，常用於臨床診斷與語言研究。然而，缺乏大規模、標準化、精準註記的中文兒童語料，使得語法分析既耗時又容易受主觀影響，現有自動化工具難以滿足臨床和研究需求。本研究提出 MINAS (Mandarin Intelligent Narrative Assessment of Syntax for Children)，結合 MAIN 故事情境與 MAPS-R 語法架構，建立涵蓋四個類別、20 個指標的中文敘事語料資料集。我們以 Prompt Engineering 評估商用模型（ChatGPT-4、Claude Sonnet 4、Gemini 2.5 Flash、DeepSeek），並以 LoRA 微調開源模型（Chinese RoBERTa、OpenHermes-2.5）。實驗結果顯示，Few-shot Prompt 能提升多數指標的辨識準確度；LoRA 微調則在名詞與動詞短語上表現更佳，但在複雜句型仍具挑戰。本研究驗證了 LLM 應用於「中文兒童敘事語料語法分類」的可行性，展現其在臨床與語言研究的潛力。

Abstract

Children's narrative ability is an important indicator of language development and is commonly used in clinical diagnosis and linguistic research. However, the lack of large-scale, standardized, and accurately annotated Chinese child language corpora makes grammatical analysis both time-consuming and prone to subjectivity, while existing automated tools fall short of clinical and research needs. This study introduces MINAS (Mandarin Intelligent Narrative Assessment of Syntax for Children), which integrates the MAIN story framework with the MAPS-R syntactic framework to construct a Chinese narrative corpus encompassing four categories and 20 indicators. We evaluated commercial models (ChatGPT-4, Claude Sonnet 4, Gemini 2.5 Flash, DeepSeek) through prompt engineering, and fine-tuned open-source models (Chinese RoBERTa, OpenHermes-2.5) with LoRA. Experimental results show that few-shot prompting achieves high accuracy across most indicators, while fine-tuning with LoRA achieves better performance in noun and verb phrase identification but is not as good for complex sentence structures. This study validates the feasibility of applying large language models to syntactic classification of Chinese child narrative corpora, highlighting their potential in clinical applications and linguistic research.

關鍵字：兒童語言評估、語法分類、大型語言模型、少樣本學習

Keywords: Child Language Assessment; Syntactic Classification; Large Language Models; Few-shot Learning

1 Introduction

兒童語言能力的發展是語言學與語言病理學的重要研究議題。特別是在語言學習初期，敘事能力(narrative ability)被視為整合語音、語法、語意與篇章組織的綜合指標，能夠有效評估兒童的語言表達與理解發展(Berman et al., 1994)。近年研究指出，兒童的敘事結構能力與其語言障礙、語用能力與認知表現息息相關，因此也逐漸成為臨床診斷的重要依據。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 184–192
November 20-22, 2025 ©2025 Association for Computational Linguistics

為此，MAIN（Multilingual Assessment Instrument for Narrative）與 MAPS-R（Multidimensional Assessment of Preschool Syntax – Revised）等工具陸續被提出。MAIN 透過故事圖片刺激，提供標準化的敘事引導情境 (Gagarina et al., 2012)，已應用於全球 65、90 餘語言；MAPS-R 則是著重於針對兒童語法能力進行詳細的系統評估 (Cheung et al., 2024)，從名詞短語、動詞短語、介詞短語到句型結構等 20 項語法分類，為中文語言臨床工作者提供了評估的參考架構。然而，即使評估架構逐漸成熟，但人工標註成本高、資料量有限、中文語料的採樣方法、內容與標註格式差異極大，都導致自動化系統難以精準建立與驗證。

近年來，大型語言模型（Large Language Models, LLMs）如 ChatGPT-4、Claude Sonnet 4、Gemini 2.5 Flash 與 DeekSeek 迅速發展，其語言理解與分類能力已在多項自然語言任務中達到極佳表現。利用提示工程（Prompt Engineering）、少樣本學習（Few-shot Learning）、模型微調（Fine-tune model）等技術，研究者能在不需要大量標註資料的前提下，引導模型完成複雜語言任務，包含語法結構分析、篇章分類、語意判斷等。然而，目前還未有針對「中文兒童敘事語料之語法結構分類」的 LLM 應用，亦缺乏可驗證的語料資源與方法驗證。

因此，本研究提出一套以 MAIN 故事為基礎、融合 MAPS-R 語法架構設計之中文敘事語料分類資料集，並結合提示工程與 LLM 模型進行語法指標辨識任務，針對名詞短語、動詞短語、介詞短語與句型類別等語法指標進行分析。藉由比較不同模型（如 ChatGPT-4、Claude Sonnet 4、Gemini 2.5 Flash 與 DeekSeek）與提示設計，探討其於兒童語言能力自動評估任務的應用潛力，本研究為探索中文語言臨床實務與語言學研究中的應用潛力。

2 Related Work

2.1 MAIN 工具與中文語法標註與語料資源建構

在兒童語言研究與臨床實務研究中，敘事能力的量測通常透過多種標準化工具來進行其中包含 ENNI（Edmonton Narrative Norms Instrument）、Renfrew Bus Story（RBS）、MAIN。

ENNI(Schneider et al., n.d.)針對 4–9 歲兒童語言能力的敘事評估工具以收集語言資料並建立本地語言表現的標準樣本數據。Renfrew Bus Story 是一個透過聽故事後復述的方式，來評估兒童口語敘事能力與語言發展的標準化測驗工具。而 MAIN（Multilingual Assessment Instrument for Narratives）是近年在語言學與語言治療領域中出現的敘事能力評估工具。該評估工具利用模範故事、故事複述及自主講述為基礎，能有效的評估兒童在語意、句法或敘事結構上的表現，在多國學者合作開發下，目前現已有 90 多種語言版本，在全球 65 多個國家使用。

在華語兒童語言研究中，中文資源相較有限。國際間規模最大的兒童語料庫 CHILDES 收藏了台灣的兒童語料，但語料來自不同的研究，研究者參與誘發發言的程度不一，導致內容複雜度差異頗大，而且系統工具 CLAN 是以 1984 年該系統創立時的處理方法，以句子為單位，逐句分層(tier)標示詞類與語法關係，並且一律強制使用簡體字，形成許多使用上的困難(MacWhinney & Snow, 1985); Sinica Treebank (Huang et al., 2000)雖提供了語法標註，但以成人語料為主。因此本研究以 MAIN 為基礎，建立華語兒童敘事語句的誘發語法分類資料集，使用臨床資料做實際的驗證及測試，補足現有資源不足之處，結合生成式 AI 進行自動評估功能，快速掌握兒童敘事語法能力。

2.2 商用大型語言模型的應用

大型語言模型在自然語言處理任務上已展現高度潛力，尤其在語句分類、語法結構判斷、語意辨識等方面，已被廣泛應用於語料分析。本次研究的模型包含 ChatGPT-4、Claude Sonnet 4、Gemini 2.5 Flash 與 DeepSeek-V3。ChatGPT-4 (OpenAI et al., 2024)其架構基於多層 Transformer 編碼器，能夠處理更長的文本輸入，並且更準確地掌握句子的語意與上下文；Claude Sonnet 4 該模型具備良好的語言理解、視覺分析、電腦操作與工具使用能力，特別擅長進行複雜的程式設計與推理任務；Gemini-2.5 Flash (Comanici et al., 2025)此模型可在需要時啟動內部「思考」機制以提升理解力和計劃能力，出色處理分類、翻譯、

程式碼執行等任務；DeepSeek-V3 (DeepSeek-AI et al., 2025)該模型結合 Multi-Head Latent Attention（MLA）與 Multi-Token Prediction（MTP）技術，在數學、程式碼與知識推理任務中展現最先進表現。

2.3 開源大型語言模型的發展

LLaMA 系列(Touvron et al., 2023)以開放權重與高效能架構為特色，是具代表性的開源 LLM。LLaMA-1 提供 7B、13B、33B 與 65B 四個版本，採用 80 Transformer layers 與 64 attention heads 的組態；其後的 LLaMA-2 延續了 LLaMA-1 的架構，在語料選取與訓練策略上進行優化，釋出 7B、13B 與 70B 參數的版本，其在多項基準任務上表現卓越，並成為眾多研究與應用的基礎；Mistral 系列中的 Mistral-7B 包含 32 個 Transformer layers、32 個 attention heads，並使用 Grouped Query Attention (GQA) 及 Sliding Window Attention (SWA)機制，能有效處理長序列輸入；OpenHermes 系列是在 Mistral-7B 基座模型上進行微調的開源模型，使用了大量程式碼相關指令資料、由 GPT-4 生成的訓練樣本，以及其他 AI 領域公開語料 (Teknium/OpenHermes-2.5-Mistral-7B· Hugging Face, 2024)。此外，社群亦釋出了基於 LoRA/QLoRA（Low-Rank Adaptation）的參數效率微調版本，使其能在資源有限的環境中應用。

2.4 Prompt Engineering 及 Fine-tune

大型語言模型（Large Language Models, LLMs）的快速發展，提示工程（prompt engineering）逐漸成為了提升模型效能的重要方式。few-shot prompting (Brown et al., 2020)顯示在無需額外訓練的情況下，僅透過設計少量範例提示即可顯著改善模型表現；Chain-of-Thought (CoT)方法 (Wei et al., 2023)則透過引導模型生成中間推理步驟，提升數學與邏輯任務的正確率。本研究也使用 finetuning 調整模型的表現。相較於 prompt engineering 的低成本與靈活性，finetuning 能針對特定任務進行更穩定與精準的調整，但其缺點是需要額外的大量標註資料與計算資源(Ziegler et al., 2020)。

3 Method

本研究採用兩種架構針對中文兒童敘事語料進行語法結構分類。詳細研究流程如圖 1 所示。

i. Prompt Engineering：針對商用 LLM 使用提示工程。
ii. LLM fine-tuning：使用 LoRA(Hu et al., 2021)對開源模型進行參數微調。

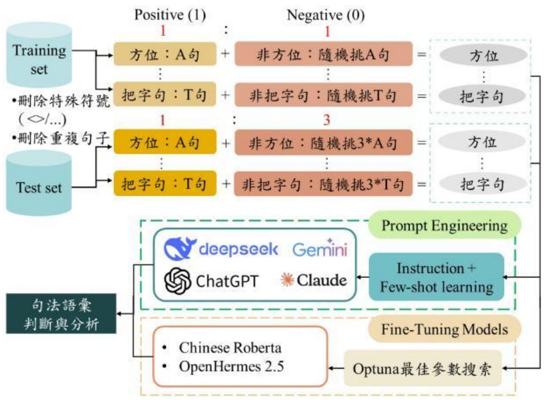


圖 1. 實驗流程圖

整體研究流程包含資料前處理、Prompt 設計、超參數最佳化、模型訓練與評估五個主要階段。首先建立基於 MAPS-R 架構的中文兒童敘事語法分類資料集，接著設計 Prompt 模板用於引導 LLM 進行語法判斷，另外也使用 Optuna (Akiba et al., 2019)進行超參數自動化搜尋，最後透過多個商用模型及開源模型的比較驗證方法的有效性。

3.1 Dataset

本研究使用的語料取自 MAPS-R 資料集 (Cheung et al., 2024)，該資料集是由三位接受 MAPS-R 編碼標準訓練的語言發展或對外進行華語教學領域的研究助理，負責中文兒童看圖敘事語料的分類。隨機抽取 20%的樣本進行 Inter-rater reliability (IRR)，計算出的 kappa 值 0.91 至 0.95。語料來源為使用 MAIN (Gagarina et al., 2012)的故事圖片引導所收集的兒童自然敘事語句。經過斷句處理與語法結構標註，能夠真實反映兒童語言發展的特徵與差異。

依據 MAPS-R 架構，兒童語言發展評估涵蓋名詞短語、動詞短語、介詞短語、句子類別等四大分類共 20 種語法指標。比如說，「他用腳踢門」符合介詞短語中的「動前介詞」指標；「拿出一本書來」符合動詞短語中的

「趨向補語」指標。

在資料前處理階段，我們移除語料中與語意無關之特殊符號、去除重複句子，確保訓練集與測試集間的資料獨立性。針對每項語法指標採用二元分類架構建立正負樣本：符合特定語法指標的句子標記為正樣本（Positive），不符合該語法指標的句子則標記為負樣本（Negative），負樣本從其他語法指標的正樣本中隨機抽取，確保樣本間的語言複雜度相當。

在樣本配置方面，訓練集採用 1:1 的正負樣本比例，測試集則採用 1:3 的正負樣本比例。各語法指標的樣本數量介於 10 至 1000 餘筆不等，反映不同語法特徵在兒童語料中的出現頻率差異。完整的資料集分布情況詳見表 1。

指標	名稱	Training set		Test set	
名詞短語		Pos	Neg	Pos	Neg
NP1	量詞-個	805	805	58	174
NP2	量詞-特定	195	195	58	174
NP3	X 的	352	352	40	120
NP4	X 的 Y	365	365	40	120
NP5	方位	510	510	43	129
動詞短句					
VP1	體貌標記	395	395	92	276
VP2	結果補語	462	462	69	207
VP3	趨向補語	252	252	47	141
VP4	情態補語	100	100	40	120
VP5	可能補語	100	100	40	120
VP6	數量補語	122	122	41	123
介詞短語					
PP1	動前介詞	100	100	48	144
PP2	動後介詞(補語)	102	102	42	126
句子					
S1	把字句	127	127	42	126
S2	被字句	100	100	40	120
S3	存現句	100	100	82	246
S4	複謂(連動/兼語)	255	255	83	249
S5	帶連詞複句	165	165	95	285
S6	緊縮複句	100	100	45	135
S7	感知/心理狀態動詞	100	100	93	279

表 1. 資料集大小

3.2 Prompt Engineering

如圖 2 所示，本研究採用經專家設計的 Prompt，引導大型語言模型進行準確的語法判斷。Prompt 模板包含兩個核心部分：

- Instruction：明確定義任務要求與輸出格式，並詳細解釋各項句法概念。為了提升模型的判斷準確性與一致性，指令中加入明確的判斷標準，規定模型須以二元分類形式（1/0）回應，並強制輸出判斷理由，以利追蹤模型推論過程與確保分析結果的可解釋性。
- Few-shot Learning： 除提供語法定義中

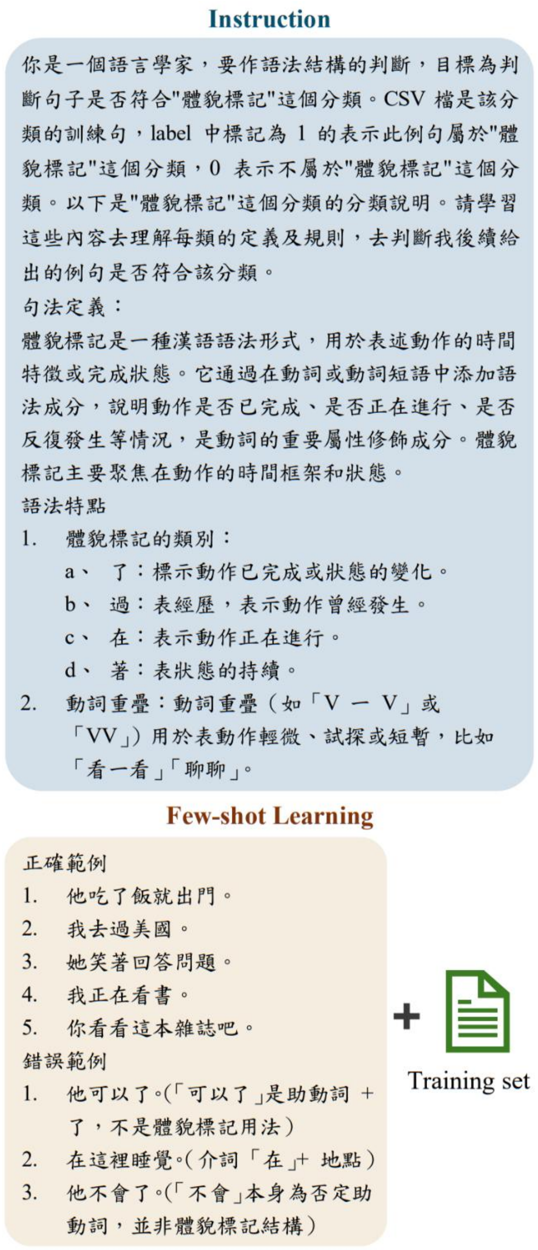


圖 2. Prrompt 模板

所附的正反範例句及錯誤分析外，進一步加入來自訓練集的句子作為示例，以增強模型對特定語法指標的判斷能力，並提升其語法概念的泛化能力。

將上述訓練集與 Prompt 輸入四種商用 LLM，分別為 ChatGPT-4、Claude Sonnet 4、Gemini 2.5 Flash 與 DeepSeek，以進行短期任務記憶與學習，隨即使用測試集進行效能評估，以驗證 LLM 在特定語法結構識別任務中的準確性。

3.3 Fine-Tuning Models

本研究使用 Chinese RoBERTa-wwm-ext (Chinese Roberta) 預訓練模型 (https://huggingface.co/hfl/chinese-roberta-wwm-ext) 和 OpenHermes-2.5-Mistral-7B-GPTQ (OpenHermes-2.5) 預訓練模型 (https://huggingface.co/TheBloke/OpenHermes-2.5-Mistral-7B-GPTQ) 進行分析評估。RoBERTa (Liu et al., 2019)是一種基於 Transformer 架構的深度學習模型，而 Chinese Roberta 進一步引入了 Whole Word Masking（WWM）(Cui et al., 2021)，在遮罩任務中針對整個詞語進行遮罩，而非單一字元，使其更具挑戰性，進而提升模型捕捉中文語義與語法關係的能力。Chinese Roberta 通過這些訓練技巧，在自然語言處理任務中顯著提升模型表現與穩健性。OpenHermes-2.5 則為大型語言模型 Mistral-7B 的 GPTQ 量化版本(Jiang et al., 2023)，也基於 Transformer 架構，具備高效微調能力，可在有限 GPU 記憶體下進行 LoRA 微調，使其能適應中文語法結構識別等任務的需求。

資料預處理：20 種指標的訓練集分別被劃分為訓練集（80%）和驗證集（20%），並確保正負類別比例分佈一致。所有語料均分別透過 Chinese Roberta 和 OpenHermes2.5 的各自 Tokenizer 轉換為模型所需的輸入格式。

LoRA：本研究採用了 LoRA 微調技術，在有限的計算資源下高效地對 Chinese Roberta 與 OpenHermes-2.5 進行 Fine-Tune。LoRA 的核心原理是在預訓練模型權重矩陣旁注入兩個 low-rank 可訓練矩陣。在訓練過程中，僅調整這兩個小矩陣的參數，而原始模型的預訓練權重保持不變。此方法顯著減少了可訓練參數的數量，大幅降低記憶體消耗與訓練時間，同時有效降低在小型資料集上發生 Over-fitting。

Optuna：在參數搜索方面，採用 Optuna 框架進行超參數自動化搜尋，以確保模型在不同資料集上均能達到最佳效能，其核心優勢在於能根據過去試驗的結果，決定下一組要嘗試的參數，從而更有效率地找到最佳參數組合。

設置 Optuna 搜尋以下幾個關鍵超參數：

- learning_rate：在 1e-6 到 1e-3 的對數尺度間搜尋。
- batch_size：從[4, 8, 16, 32]中選擇。
- lora_r：從[8, 12, 16, 20, 24]中選擇。
- lora_alpha：從[4, 8, 16, 32, 64]中選擇。
- lora_dropout：在 0.0 到 0.5 間搜尋，步長為 0.1。

每種資料集皆獨立進行 30 次試驗，以驗證集上的 validation loss 作為最佳化的目標。最終，我們將每種資料集所獲得的最佳參數組合記錄下來，並用於後續的最終模型訓練。

訓練過程使用 Hugging Face 的 Trainer 類別進行，並使用 Optuna 找到的最佳超參數。為了防止 Over-fitting，我們採用 Early Stopping 機制，當 validation loss 連續 15 個 epoch 沒有改善時，訓練會自動停止，並載入表現最佳的模型權重。最終，我們使用測試集評估模型在特定語法結構識別任務中的效能。

針對 Chinese RoBERTa 的整體 DNN 架構流程詳見圖 3。輸入的語料會先經由 Tokenizer 轉換為數位格式，接著進入 Chinese RoBERTa 並使用 LoRA 進行參數微調。最終連結至 Dense Layer 與 Softmax 以進行二元分類，得到 1 或 0 的結果，分別代表 Positive 或 Negative 標籤。

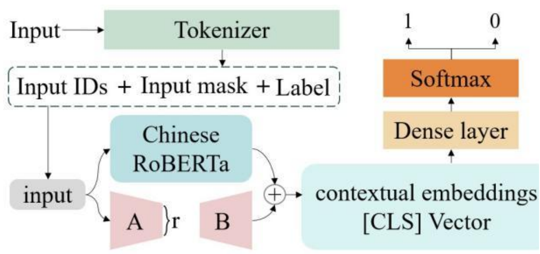


圖 3. 基於 RoBERTa-LoRA 的

文本分類 DNN 架構

4 Result and Discussion

指標	Few-shot				Zero-shot			
	Gemini	Claude	ChatGPT	Deepseek	Gemini	Claude	ChatGPT	Deepseek
NP1	**0.983**	0.966	0.975	0.922	0.945	0.952	**0.974**	0.758
NP2	0.779	**0.855**	0.689	0.769	0.780	0.790	0.775	**0.820**
NP3	0.976	0.988	0.987	**1.000**	0.975	0.981	0.870	**1.000**
NP4	0.987	**1.000**	0.909	0.963	0.980	**0.985**	0.867	0.935
NP5	**0.966**	0.930	0.913	0.945	0.960	0.967	0.966	**0.977**
VP1	**0.948**	0.928	0.879	0.845	0.917	0.910	**0.938**	0.738
VP2	0.763	0.789	0.838	**0.872**	0.724	0.745	0.644	**0.769**
VP3	**0.842**	0.793	0.832	0.839	0.835	0.823	0.825	**0.839**
VP4	0.988	**1.000**	0.833	0.975	0.976	0.980	0.951	**1.000**
VP5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
VP6	0.932	0.911	**0.975**	0.874	0.914	0.920	0.845	**0.953**
PP1	0.787	**0.922**	0.750	0.883	0.853	0.864	0.831	**0.949**
PP2	0.966	0.953	**0.977**	0.943	0.910	0.930	0.788	**0.989**
S1	**1.000**	0.976	0.822	0.988	**0.986**	0.985	0.988	0.977
S2	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
S3	0.747	**0.796**	0.793	0.755	0.742	0.750	**0.761**	0.720
S4	**0.964**	0.878	0.698	0.872	0.881	**0.907**	0.626	0.943
S5	**0.940**	0.931	0.581	0.935	0.569	0.580	0.540	**0.593**
S6	**0.977**	0.957	0.911	0.945	0.945	**0.956**	0.846	0.936
S7	0.883	0.690	**0.962**	0.690	0.617	**0.657**	0.583	0.633
Micro_F1	**0.902**	0.890	0.854	0.878	0.832	0.840	0.797	**0.844**
Macro_F1	**0.920**	0.912	0.865	0.901	0.875	**0.884**	0.831	0.876

表 2. Few-shot 各語法類別的 F1-Score

4.1 Prompt Engineering 效果分析

表 2 呈現 Few-shot 以及 Zero-shot prompting 在各語法類別的 F1-Score 實驗結果。整體而言，多數語法指標在 Few-shot 條件下均能達到穩定且高準確度的表現，F1-Score 在 0.95 以上；然而在結構和語義較複雜的類別（如「量詞-特定」、「結果補語」、「動前介詞」及「存現句」）上，模型仍存在辨識困難，F1-Score 落在 0.792 和 0.855 之間。

在 Zero-shot 實驗結果中，雖然部分語法指標，如「X 的」、「被字句」仍能維持較高分數，但在多數類別上，模型表現均低於 Few-shot 條件，尤其在句法結構與語義較為複雜的類別中差距更為明顯，例如「感知/心理狀態動詞」類別在 Few-shot 條件下的 F1-Score 較 Zero-shot 提升 30.5%，而帶連詞複句則提升 34.7%。

由此可見，單純依賴模型的內建知識和語法定義並不足以應對複雜的語法任務，而 Few-shot 的引入則能顯著提升模型在語法指標識別上的整體表現。Zero-shot 雖展現出模型固有語言知識的潛力，但在多數類別上仍表現不足，尤其是結構與語義多樣化的句型中。相比之下，Few-shot 透過少量範例有效縮小了模型的判斷偏差，顯示出在資源有限的情境下，Prompt 設計仍是提升大型語言模型語法處理能力的重要策略。

4.2 Fine-tuning 效果分析

表 3 列出兩個 LLM 在 20 個語法指標的最終的 F1-Score，整體而言，兩個模型在名詞短語與簡單動詞短語上均能達到高 F1-Score（0.95 – 1.00），且所需訓練迭代較少；相較之下，介詞短語及複雜句型（如「複謂」、「帶連詞複句」、「緊縮複句」、「感知/心理狀態動詞」）的 F1-Score 較低。

比較兩個模型可見，Chinese Roberta 在大部分語法類別上的 F1-Score 稍高於 OpenHermes 2.5。OpenHermes 2.5 部分動詞短語與句子類別的 F1-Score 與 Chinese Roberta 相近，但在語義和結構複雜的項目仍稍差。

綜合來看，Fine-tune 在大多數語法指標的識別上也能展現出不錯的效果，Macro_F1-Score 達到 0.854。對大部份名詞與動詞短語的辨別能力尤為顯著；然而對結構或語義複雜的句子，模型仍存在一定限制，未來可針對

這些類型強化資料基礎或探索更精細的 Fine-tuning 策略。

指標	Chinese Roberta	OpenHermes 2.5
NP1	**0.983**	0.953
NP2	**0.836**	0.798
NP3	**1.000**	0.982
NP4	**0.987**	0.964
NP5	**0.743**	0.697
VP1	**0.941**	0.876
VP2	**0.841**	0.783
VP3	**0.729**	0.681
VP4	**0.889**	0.824
VP5	**0.930**	0.851
VP6	**0.795**	0.746
PP1	**0.822**	0.783
PP2	**0.848**	0.802
S1	**0.977**	0.917
S2	**1.000**	0.924
S3	**0.766**	0.546
S4	**0.783**	0.698
S5	**0.891**	0.745
S6	**0.750**	0.678
S7	0.547	**0.603**
Micro_F1	**0.828**	0.774
Macro_F1	**0.854**	0.793

表 3. Fine-tune 各語法類別的 F1-Score

5 Error Analysis

為進一步理解模型在語法指標上的判斷偏差，本研究根據 Macro_F1-Score 表現最佳的 Gemini 在部分指標上分數相對較低的情況，選取量詞-特定、結果補語、動前介詞及存現句進行錯誤分析，檢視 False Positive（FP）與 False Negative（FN）案例，並比較模型判斷邏輯與應有標註的差異。

5.1 量詞-特定：

「帶特定量詞」的名詞短語用於表達對事物的特定量化，通過量詞與名詞搭配增加語義精確性，例如「一瓶水」、「一碗飯」。FP 案例如：「這份報告轉交給部門經理了」、「我跑了五分鐘就累了」、「這幾件都很好看，我想要亮的」。模型將「份」、「分鐘」等誤判為特定量詞修飾名詞，但實際上應視為一般個體量詞或度量單位，非真正特定量詞。

5.2 結果補語

結果補語表示動作完成後的結果，例如「我把衣服洗乾淨了」。FN 案例如：「牠跌倒」，模型識別「跌倒」是一個複合動詞，卻未識別出其結果補語為「倒」。FP 案例如：「因為聽不清楚，所以我又問了一次」，模型將「聽不清楚」判為結果補語，實際上應屬情態補語，結果補語的否定形式需在主要動詞前加「沒有」。

5.3 動前介詞

動前介詞出現在動詞前，對句子的謂語提供輔助資訊，用於修飾或限定動作的條件、時間、地點、對象等。FP 案例如：「我擠到人群裡面了」、「她塞了一個蘋果到背包裡」。模型將「到人群裡面」、「到背包裡」誤判為動前介詞，但語法上介詞出現在動詞後應屬動後介詞。

5.4 存現句

存現句用於陳述物品存在或事件發生，例如「桌子上有一本書」。FN 案例如：「警察局需要有槍」、「我這裡有三塊」，模型將強調必要性或擁有的句子誤判為非存現句。FP 案例如：「還有吐司」、「他還有一個翅膀」，模型識別為存現句，但實際「還有」表示在已有基礎上額外加上，並非純粹存在陳述。

5.5 Fine-tuning 的挑戰與限制

綜合本研究實驗結果，Fine-tuning 難以在相同資料條件下穩定超越 Prompt Engineering，可能的原因如下：

- 由於大型 LLM 的訓練語料龐大且涵蓋範圍廣泛，模型在預訓練過程中往往已具備一定的語言規則與推理能力。能在規則性較強的任務（如句法判斷）快速展現適應性。相較之下，Fine-tuning 資料量不夠充分時，模型可能無法有效收斂或容易過擬合，導致特定訓練句數較少的語法指標表現不佳。
- Prompt Engineering 效果穩定：透過在 Prompt 中提供適當的任務背景和範例，模型通常能夠準確地進行分類。然而，微調模型在測試語料較為罕見或與訓練集的句型差異較大的資料集上，表現可能不如預期。

- 模型成效與超參數配置的關聯：Fine-tuning 的成效依賴於學習率、batch size、epoch 與正則化等超參數設定，即使使用 LoRA 也可能因調校不足而難以達到參數收斂至最佳效果。

6 Conclusion

本研究基於 MAIN 故事情境 MAPS-R 的中文兒童敘事語料和語法架構，提出 MINAS 系統，並結合 Prompt Engineering 與 Fine-tuning 策略進行語法結構辨識。實驗結果顯示，Gemini 搭配 prompt engineering 可達 0.902 的 Micro_F1 與 0.920 的 Macro_F1，其中在「特定量詞」、「結果補語」、「動前介詞」等語法指標上的誤判較為明顯，但對多數語法指標都有較好的辨識準確度(F1 > 0.9)。LLM Fine-tuning 整體準確率稍差，Roberta 達到 0.828 的 Micro_F1 與 0.854 的 Macro_F1，其在名詞與動詞短語的分類上表現卓越，F1-Score 可達 0.95-1.00，證明其在處理特定語法任務時的有效性。然而，對於結構與語義複雜的句型，如「結果補語」和「存現句」，Fine-tuning 模型仍存在誤判。這反映了中文兒童語料在語法與語義表達上的多樣性，也提示 Fine-tuning 模型在測試集中較少見或與訓練集差異明顯的句型上可能表現受限。

此外，本研究仍存在若干限制：資料集規模較小且語法指標分佈不均，可能影響模型的泛化能力；然而，這樣的分佈特性亦真實反映了兒童自然語料中各類語法結構的實際出現頻率差異。傳統句法分析模型未被納入比較，主要原因在於本研究的多項語法指標同時涉及語義判斷（例如趨向補語、存現句等），而傳統句法模型主要聚焦於結構層面的分析，難以處理語義層面的判斷，因此未被納入主實驗比較。

本研究嘗試在中文兒童語料上探索大語言模型的語法與語義理解能力，驗證大型語言模型在中文兒童敘事語法分類任務上的可行性與應用潛力。實驗結果證實使用 LLM 進行中文兒童語法分類的可行性，並為語言臨床評估與語言學研究提供自動化分析數據。未來，我們將持續蒐集臨床語料、擴充語料集大小，並探索更精細的微調策略，以進一步提升大型語言模型在兒童敘事能力分析的可行性與判斷準確率。

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework* (No. arXiv:1907.10902). arXiv. https://doi.org/10.48550/arXiv.1907.10902

Berman, R. A., Slobin, D. I., Aksu-Koç, A. A., Bamberg, M., Dasinger, L., Marchman, V., Neeman, Y., Rodkin, P. C., Sebastián, E., & et al. (1994). *Relating events in narrative: A crosslinguistic developmental study* (pp. xiv, 748). Lawrence Erlbaum Associates, Inc.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (No. arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Cheung H., Ch L., & Cj C. (2024). Measuring productive syntactic abilities in Mandarin-speaking children in Taiwan. *Clinical linguistics & phonetics*, *38*(11). https://doi.org/10.1080/02699206.2024.2302549

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., … Bhumihar, N. K. (2025). *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities* (No. arXiv:2507.06261). arXiv. https://doi.org/10.48550/arXiv.2507.06261

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-Training with Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3504–3514. https://doi.org/10.1109/TASLP.2021.3124365

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., … Pan, Z. (2025). *DeepSeek-V3 Technical Report* (No. arXiv:2412.19437). arXiv. https://doi.org/10.48550/arXiv.2412.19437

Gagarina, N. V., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., Bohnacker, U., & Walters, J. (2012). MAIN: Multilingual assessment instrument for narratives. *ZAS Papers in Linguistics*, *56*, 155–155. https://doi.org/10.21248/zaspil.56.2019.414

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models* (No. arXiv:2106.09685). arXiv. https://doi.org/10.48550/arXiv.2106.09685

Huang, C.-R., Chen, F.-Y., Chen, K.-J., Gao, Z., & Chen, K.-Y. (2000). Sinica Treebank: Design criteria, annotation guidelines, and on-line interface. *Proceedings of the Second Workshop on Chinese Language Processing: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 12*, 29–37. https://doi.org/10.3115/1117769.1117775

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B* (No. arXiv:2310.06825). arXiv. https://doi.org/10.48550/arXiv.2310.06825

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (No. arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*(2), 271–295. https://doi.org/10.1017/S0305000900006449

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (No. arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Plante, E. (n.d.). *The diagnostic and predictive validity of the Renfrew Bus Story*. Retrieved September 10, 2025, from https://www.academia.edu/13384934/The_diagnostic_and_predictive_validity_of_the_Renfrew_Bus_Story

Schneider, P., Hayward, D., & Dubé, R. V. (n.d.). *Évaluer grâce au « Edmonton Narrative Norms Instrument » une histoire contée à partir d'images Storytelling from pictures using the Edmonton Narrative Norms Instrument*.

Teknium/OpenHermes-2.5-Mistral-7B · Hugging Face. (2024, April 15). https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (No. arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (No. arXiv:2201.11903). arXiv. https://doi.org/10.48550/arXiv.2201.11903

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2020). *Fine-Tuning Language Models from Human Preferences* (No. arXiv:1909.08593). arXiv. https://doi.org/10.48550/arXiv.1909.08593

LOBSTER🦞: Linguistics Olympiad Benchmark for Structured Evaluation on Reasoning

Da-Chen Lian
Graduate Institute of Linguistics
d08944019@ntu.edu.tw

Ri-Sheng Huang
Dept. of CSIE
r13922102@csie.ntu.edu.tw

Pin-Er Chen
Graduate Institute of Linguistics
f10142001@ntu.edu.tw

Chunki Lim
Graduate Institute of Linguistics
r14142001@ntu.edu.tw

You-Kuan Lin
Dept. of Elec. Engineering
conlangtaiwan@gmail.com

Guan-Yu Tseng
Graduate Institute of Linguistics
r14142007@ntu.edu.tw

Zhen-Yu Lin
Dept. of FLL
a0985026048@gmail.com

Pin-Cheng Chen
Dept. of FLL
b10102102@ntu.edu.tw

Shu-Kai Hsieh
Graduate Institute of Linguistics
shukaihsieh@ntu.edu.tw

National Taiwan University

Abstract

We propose the Linguistics Olympiad Benchmark for Structured Evaluation on Reasoning, or LOBSTER🦞, a linguistically-informed benchmark designed to evaluate large language models (LLMs) on complex linguistic puzzles of the International Linguistics Olympiad (IOL). Unlike prior benchmarks that focus solely on final answer accuracy, our benchmark provides concrete evaluation protocols and rich typological metadata across over 90 low-resource and cross-cultural languages alongside the puzzles. Through systematic evaluations of state-of-the-art models on multilingual abilities, we demonstrate that LLMs struggle with low-resource languages, underscoring the need for such a benchmark. Experiments with various models on our benchmark showed that IOL problems remain a challenging task for reasoning models, though there are ways to enhance the performance—for example, iterative reasoning outperforms single-pass approaches in both final answers and explanations. Our benchmark offers a comprehensive foundation for advancing linguistically grounded, culturally informed, and cognitively plausible reasoning in LLMs. [1]

Keywords: reasoning, large language model, benchmark, linguistics olympiad

[1]The benchmark and the source code can be found at https://github.com/lopentu/LOBSTER.

1 Introduction

While advances in LLM have revolutionized natural language processing, significant challenges persist in achieving robust reasoning capabilities—particularly for tasks requiring multi-step abstraction, symbolic verification, and constraint-based hypothesis testing. Several reasoning-enhancement paradigms have emerged with the hope to solve more complex problems, such as hybrid tool-integrated approaches (He et al., 2025; Gao et al., 2025; Paranjape et al., 2023; Schick et al., 2023; Wu et al., 2025), or agentic systems (Li et al., 2025; Ke et al., 2025).

The International Linguistics Olympiad (hereinafter abbreviated as IOL; 2003-2025) presents uniquely challenging problems that require solvers to induce linguistic rules from micro-data, often in low-resource or unfamiliar languages. These problems test not just surface-level pattern recognition, but demand multi-step abstraction, structural reasoning, and cultural inference. Comprising four parts (see Appendix A.1), an IOL problem is meticulously crafted to be self-contained, without the necessity of any prior knowledge in linguistic rules. The logical consistency and sufficiency thus allows participants to decode the underlying linguistic rules purely through reasoning and pattern analysis (Bozhanov and Derzhanski, 2013), the low-resource nature of the languages in which these problems made offers an isolated envi-

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 193–229
November 20-22, 2025

ronment to test the reasoning performance of models. (See Section 3)

In addition to abstract linguistic reasoning, some IOL problems incorporate elements that go beyond standard textual input, requiring models to process non-standard scripts, phonetic transcriptions, or visual symbol systems such as maps or family trees. Some problems involve rare or extinct writing systems—occasionally ones not yet fully encoded in Unicode—demanding the recognition and manipulation of unfamiliar glyphs (Shih et al., 2025). Others rely on International Phonetic Alphabet (IPA) representations, tone contour symbols, or constructed orthographies that encode morphophonemic information. A subset of tasks also includes pictographic cues, spatial arrangements, or logical diagrams (see Appendix A.2), which are essential to its decipherment. While recent vision-language models have made progress in visual and text input jointly, their ability to integrate these modalities with complex reasoning remains limited.

Another distinctive aspect of IOL problems lies in their **cross-cultural and semantic depth**. Beyond the structural reasoning over phonology, morphology, and syntax, many problems explicitly involve semantic inference, cultural conceptualization, or sociolinguistic reasoning—for instance, deciphering kinship terms, numeral systems, metaphorical extensions, or culturally situated deixes. These tasks compel both human and AI solvers to imagine how meaning might be constructed in unfamiliar cultural worlds, often requiring *cross-linguistic abstraction* or *anthropological imagination*. For LLMs, this poses a profound challenge: it tests their ability to generalize across not only linguistic structures but also cognitive and cultural domains. IOL problems, therefore, serve not only as puzzles of language form but as tests of situated meaning-making and cultural flexibility, offering a rigorous probe into the limits of LLMs' representational and interpretive capacity across diverse human experiences.

These complex challenges expose the limitations of current LLMs and existing evaluation methods, which often prioritize final-answer accuracy over the reasoning process.

2 Review of Past Studies

Reasoning models and reason-enhancing paradigms enable LLMs to actively explore solutions, rather than just passively generate text. Their efficiency is frequently evaluated through human-level reasoning benchmarks like the International Linguistics Olympiad (IOL) (Şahin et al., 2020; Chi et al., 2024), where success requires inferring linguistic structures from constrained datasets, mirroring real-world challenges in rule abstraction, cross-linguistic generalization, and constraint satisfaction.

2.1 Reasoning on Linguistic Structures

Reasoning on linguistic structures presents unique challenges, when compared to other reasoning domains such as math or coding. Unlike purely symbolic systems, understanding human languages requires world knowledge, cultural context, and common sense. For example, the word for "five" and "hand" is the same in some languages because there are five fingers on a hand. This requires the model to also infer of a semantical link between the two senses; it is inconceivable from a symbolic inductive logical perspective.

For the classic Rosetta Stone problems,[2] the inference task is in a sense a more complex variant of the "infer one form of a word/phrase/sentence to another" task.

The induction task has long been of interest to linguists (Durham and Rogers, 1969), as it mirrors what linguists do in a field study. This induction task has been framed in at least two ways. One perspective treats it as a program synthesis problem, where the goal is to generate a "program"—a set of formal rules—that transforms inputs to outputs (Naik et al., 2024). This has led to the development of domain-specific languages for expressing such string transformations (Vaduguru et al., 2021). Alternatively, the task can be viewed as constrained text generation, where specialized architectures are designed to model linguistic phenomena (Lu et al., 2024).

A complementary line of research explores augmenting LLMs with explicit linguistic knowledge. Rather than relying solely on induction from examples, this approach provides models with resources like dictionaries, morphological analyzers, or grammar books, mimicking how a human linguist might consult reference materials (Zhang et al., 2024). While the ability to leverage such

[2]Given a set of sentences in an unknown language and their corresponding translations, the agent should infer the underlying rules, such as grammar, meaning of each word, or spelling changes in the unknown language.

grammatical descriptions can be systematically evaluated (Tanzer et al., 2024), their utility is task-dependent: for translation, performance gains stem from parallel examples rather than grammatical explanations, which are better suited for targeted linguistic analysis tasks (Aycock et al., 2025). Such nuances call for more research on the intersection of LLMs and linguistics expertise.

2.2 Relevant Benchmarks from Linguistics Olympiads

To evaluate the capabilities of LLMs on complex reasoning tasks, researchers have developed various benchmarks. The following are some benchmarks relevant to Linguistics Olympiad problems:

- **LingOly** (Bean et al., 2024):[3] With 1,133 linguistic puzzles from the UK Linguistics Olympiad (UKLO),[4] it excludes image-based puzzles, non-Latin scripts, and open-ended questions to ensure machine-scorability. The evaluation is exact-matched, excluding fuzzy matches and normalizing Unicode variations, to ensure linguistic precision. Less strict metrics like ROUGE and BLEU were analyzed, but the primary focus remains on context-dependent reasoning.

- **Linguini** (Sánchez et al., 2024):[5] This benchmark also extracted data from IOL problems, covering low-resource languages and three core task types: sequence transduction , fill-in-the-blanks, and number transliteration (i.e. digit-to-text conversion). The evaluation uses exact match accuracy and the softer chrF metric to assess performance on structured linguistic inference.

- **IOLBENCH** (Goyal and Dan, 2025):[6] 90 of the IOL Problems were digitalized into text or structured representation through LLM-based document recognition, including some multimodal components. While it take cares of free-response answers through different grading metrics, the LLM-based unverified data construction made most of the problem in the dataset ill-formed.

[3]Relevant resources for LingOly can be found on GitHub: https://github.com/am-bean/lingOly.

[4]https://www.uklo.org/

[5]Relevant resources for Linguini can be found on GitHub: https://github.com/facebookresearch/linguini

[6]Relevant resources for IOLBENCH can be found on GitHub: https://github.com/Satgoy152/ling_llm

Existing benchmarks for IOL-style tasks have demonstrated the promising capabilities of LLMs in handling complex linguistic reasoning. However, several critical limitations remain that constrain both fine-grained evaluation and meaningful model improvement.

First, most current evaluations rely predominantly on exact-match accuracy of the final answers, without considering the plausibility, internal consistency, rules used to explain the answers, or are logical coherence of intermediate reasoning steps. This narrow focus obscures whether models are genuinely applying linguistic principles or merely relying on pattern recognition and heuristic guessing. Such a limitation hampers our ability to diagnose reasoning failures and systematically improve model understanding.

Specifically, these methods often (i) lack rigorous alignment with linguistic knowledge bases, (ii) fail to capture the reflective, iterative, and self-corrective nature of human linguistic reasoning, and (iii) inadequately represent the hierarchical and multi-layered reasoning structures characteristic of IOL challenges. As a result, existing evaluation paradigms are insufficient for capturing the depth, correctness, and explanatory richness of linguistic problem-solving processes. This highlights the need for more sophisticated evaluation methodologies specifically tailored for linguistic reasoning contexts.

3 Motivation: Probing the Limits of LLMs

As Joshi et al. (2020) highlight, the vast majority of the world's languages are low-resource, and their unique linguistic features are underrepresented in pre-training corpora. This skew towards high-resource languages like English hinders model performance and the potential for cross-lingual transfer, even for typologically similar languages (Pires et al., 2019).

To empirically ground the need for a more nuanced evaluation benchmark, we assessed a state-of-the-art model, Gemini-2.5-flash, on a multilingual translation task using the FLORES-200 dataset (NLLB Team et al., 2022). Our experiment, which covered 204 languages, revealed critical limitations (see Appendix I for full details). We found that:

1. Performance is strongly correlated with resource availability. The model frequently

failed to generate any output for the lowest-resource languages (Class 0).

2. A significant performance asymmetry exists based on translation direction. The model performed substantially worse when translating from English *to* a target language ($E \rightarrow T$) than in the reverse direction ($T \rightarrow E$), especially for low-resource languages.

3. Statistical analysis confirmed that language family and resource class are highly significant predictors of translation quality, while script was not.

These findings demonstrate that even powerful models struggle with genuine multilingual tasks, often failing at the basic level of text generation for a large portion of the world's languages. This underscores the inadequacy of benchmarks that focus only on high-resource languages or overlook reasoning failures, motivating our development of LOBSTER.

4 LOBSTER: Linguistics Olympiad Benchmark for Structured Evaluation on Reasoning

The IOL problems exhibits a wide range of typological diversity, an essential step in understanding the nature of such a benchmark in profiling the distribution of languages, for which existing LLM benchmarks rarely account. Regarding language family, the most common language families are North American, Austronesian, Indo-European, and African (see Appendix D for the language family distribution). However, There remains a gap in understanding how models perform across different language families and typological features.

LOBSTER is built on a curated selection of past IOL problems. Unlike prior datasets, it includes enriched metadata that allows for deeper linguistic diagnostics and reasoning trace comparison. Our benchmark is intended to support: (i) accurate transcription of contents of IOL problems; (ii) typologically grounded performance analysis; and (iii) assessment of models' cross-cultural and cross-linguistic inference abilities.

4.1 Data Construction

Our benchmark consists of 96 problems (225 subproblems) sourced from the IOL archive (2003–2024). For kinship problems[7] involving family trees, we convert the graphical representations into textual relationship descriptions (see Appendix A.3 for an example of a kinship problem). We exclude problems that fully rely on image-based information or untranscribable symbols.

Since most IOL problems provide only the final solutions along with some grammatical rules, without including detailed reasoning steps, we use Gemini-2.5-pro to generate structured step-by-step solutions as gold-standard references in the benchmark. The LLM is prompted to act as a linguistics expert, producing logical deductions, linguistic rules, and problem-solving strategies that lead to the official solutions (see Appendix B for the prompt template). To ensure reliability, seven human experts and three IOL contestants manually verify and refine these reasoning chains, resolving any inconsistencies to ensure alignment with the official IOL solutions.

In summary, for each IOL problem in our benchmark, we include the transcribed problem text, the official solution, and the expert-verified, refined, LLM-generated reasoning. The latter is not used for grading but serves as a qualitative reference for human-understandable reasoning processes.

4.2 Typological Annotation

In addition, each problem within LOBSTER is annotated along multiple linguistic dimensions to facilitate a structured analysis of model performance. The current typological and problem-oriented schema is an adaptation of the UKLO classification framework[8] with the annotation being carried out by seven linguistic experts. We annotate three categories for each problem: Subject, Type, and Theme; the respective tags are detailed below, while the descriptions of each tag are shown in Appendix C.1. Also, the Glottocode is included (Hammarström et al., 2024) for each problem. Table 2 shows an example of annotations for one problem.

The distribution charts of each typological category in our benchmark are shown in Appendix D. Key findings include:

Subject and Type Distribution: Referring to Appendix E, the data suggests that Syntax and

[7] Kinship problems focus on understanding how different languages and cultures describe family relationships and naming systems.

[8] `https://www.uklo.org/technical-information/`

Morphology are the most prominent subjects in IOL problems, with Rosetta type problems being heavily focused in these areas (i.e., 17.4% and 16.5%). Semantics are distributed across multiple problem types (0.9%, 6.4%, 3.7%, 0.9%, 7.3%) compared to others. Overall, the uneven distribution implies that certin problem types are strongly associated with particular subjects (e.g., Phonology has a spike (13.8%) in Pattern type problems), while others are more diffuse.

Subject and Language Family Distribution: North American languages have the highest number of problems (14), followed by Austronesian (11), Indo-European (10), and African (10). As shown in Appendix F, Syntax is the most widely represented subject, appearing in 7 out of the top 10 language families, with the highest concentration (6.2%) in African. Morphology is the second most frequent, appearing in 9 out of the top 10 families, with multiple mid-range values (2.5%–5.0%). While Phonology stands out in Indo-European and North American, Semantics is more broadly distributed, with Austronesian, African, Australian, and Niger-Congo all having moderate percentages (around 2.5%). In summary, Syntax, Morphology, and Phonology dominate the subject distribution, with North American, Austronesian, Indo-European, and African languages showing the richest variety of subjects. More details are shown in Figures (a) and (d) in Appendix D.

Type and Language Family Distribution: Regarding Appendix G, *Match-up* problems are more common in Austronesian and North American language families. *Pattern* problems are particularly prevalent in Indo-European languages. *Rosetta* problems are the most common overall (44 problems), appearing across various language families, with especially high occurrences in African and North American languages. More details are shown in Figure 8 (b) and (d) in Appendix D.

These findings reinforce the relevance of typological and reasoning-aware annotations. They also highlight the inadequacy of answer-only metrics in capturing the richness of linguistic cognition demanded by IOL problems.

4.3 Evaluation Protocol and Metrics

Existing IOL-styled benchmarks (Bean et al., 2024; Sánchez et al., 2024; Goyal and Dan, 2025) tend to rely on exact string matching for accuracy, which fails to award partial credit for complex problems. Grading IOL solutions is rather complex and flexible. Generally, the final answer is not the sole contributor to the final score; the explanation of grammatical rules is just as important. We hence evaluate the final solution generated by the model with respect to the rules provided in official solutions.

4.3.1 Evaluation of the Final Solution

First, we assess the model-generated final solution based on two distinct components: the **answer** and the **explanation of rules**.

The **answer** refers to all the questions inside the problem, which the contestant would be asked to answer. For example, the sample problem in Appendix A.1 contains 9 questions (1 in subproblem (a), 3 in (b), and 5 in (c)). Most of the questions, such as short sentence translations, can be graded with simple string matching, but an exact match metric would be unsatisfactory in many cases. Examples include semantics problems where any synonym should be counted as correct if the term is inferred, but not copied from the problem; or questions that ask for an explanation to a certain linguistic phenomenon (not to be confused with the "explanation" part of the solution below). In these cases, various metrics can be applied (e.g., BLEU, sentence embedding) depending on the preferences of the user of our benchmark.

On the other hand, the **explanation** requires the model to write down the linguistic rules it inferred from the problem data. The official IOL problem sheet explicitly states, "Your answers must be well-supported by argument. Even a perfectly correct answer will be given a low score unless accompanied by an explanation", but the official grading rubrics are not publicly available, thus evaluating the quality of these free-text explanations poses a significant challenge unaddressed by past works. We address it with a two-stage procedure: Through **rule composition**, we convert the official solution into a discrete set of key linguistic rules, creating a gold-standard "rule checklist." We then employ an LLM grader, specifically Gemini-2.5-flash-lite, in the process for **checklist grading**. The grader is prompted to compare the model's generated explanation against our rule checklist and determine the number of gold-standard rules that were correctly described. By grading with a checklist rather than the official, free-form solution, we reduce subjectivity in the grading criteria, and minimize potential biases (e.g., self-preference) from the LLM grader.

This approach enables a stable, fine-grained, and quantitative assessment of the explanation's quality. The total score for the final solution is a weighted combination of the scores of "answer" and "explanation of rules." By default, we assign equal weight (50/50) to each component, with points distributed evenly across all subproblems for the answer and all identified rules for the explanation. With additional scores granted to the explanation, the benchmark we propose can show whether the model answers through reasoning within the problem data or through other external confounders.

5 Testing LOBSTER🦞 on Different Systems

In the previous sections, the multilingual abilities of LLMs are shown to be inadequate. Therefore, when attempting to solve an IOL problem, LLMs may not solely rely on prior knowledge about the target language or typology. To pinpoint the ability of state-of-the-art models on IOL problems, we examined a range of models on LOBSTER🦞, and verified that IOL problems pose a challenge even for state-of-the-art reasoning models.

5.1 Setup

A set of experiments was conducted using the most powerful models within budget. In addition to directly prompting the models, we also tested with various settings for the same model. To ensure numerical stability, for each problem in each setting, we obtained 5 samples and averaged over the scores. The settings include:

- **Vanilla baseline**: A direct, single-pass call to an LLM to solve the problem, following the required output format. We used OpenAI-o4-mini, Gemini-2.5-pro, and GPT-5 for the experiments, with temperature set to 0.75.

- **Guided prompt**: A major drawback of the vanilla prompting is that, usually the LLM is not familiar with the underlying assumptions of Linguistics Puzzles (e.g., "All the questions are self-contained", "The final solution should be able to explain 100% of the examples, not just 90%"). To inform the model about such nuances, we include the Introduction chapter of the book *Linguistics Olympiad: Training guide* (Neacşu, 2024) in the system prompt. As an introductory text about linguistics problem, the chapter describes the format and classification of a linguistics problem, guidelines for solution writing, and some toy examples.

- **Grammar agent**: Past work has shown that the model performs better when given explicit knowledge (Tanzer et al., 2024). In this setting, the model was provided with a reference grammar book of the target language. To do so, we constructed a database containing reference grammar books from publicly available resources, and manually labeled the language, with its Glottocode as metadata to facilitate search.

- **Mixture-of-Agents**: Following Mixture-of-Agents (MoA) (Wang et al., 2025), a multi-round framework is used, as depicted in Figure 1. The system consists of a customizable number of *Solver Agents* and *Aggregator Agents*. The idea is that iteratively collecting multiple proposed solutions may improve performance. In our setup, we used 2 agents for each layer (N=2 following the notations in the figure)—Gemini-2.5-pro and OpenAI-o4-mini. The solutions are iterated for at most 6 rounds (i.e., M=2, R $\in \{0, 1, 2, 3, 4\}$), with the last round being the "final aggregator" in the figure.

- **Single agent, multi-rounds**: Equivalent to the Mixture-of-Agent setting with N=M=1, the solution of a solver is fed into itself for multiple rounds. This setting disentangles the effect of parallel generation from iterative refinement.

5.2 Results and Analysis

5.2.1 Comparison between Models

The results are summarized in Figure 2. Based on the evaluation methods detailed in Section 4.3.1, the answer and the explanation scores are calculated separately, and a combined score ("total score") is also provided. An overview shows that the scores for the "answer" and the "explanation" are positively correlated (r=0.501). (See Appendix O)

Regarding the base models, our experiments are mainly comparing models based on Gemini-2.5-

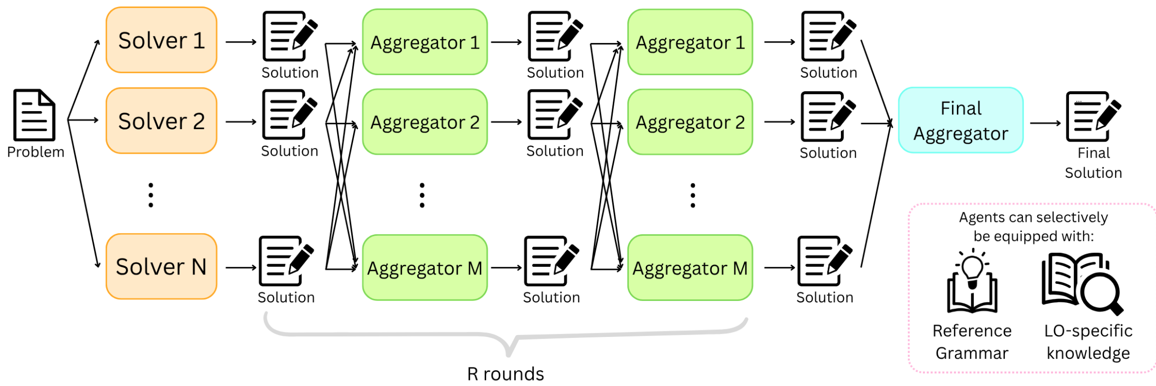


Figure 1: Multi-Agent Framework for Solving Linguistics Olympiad Problems

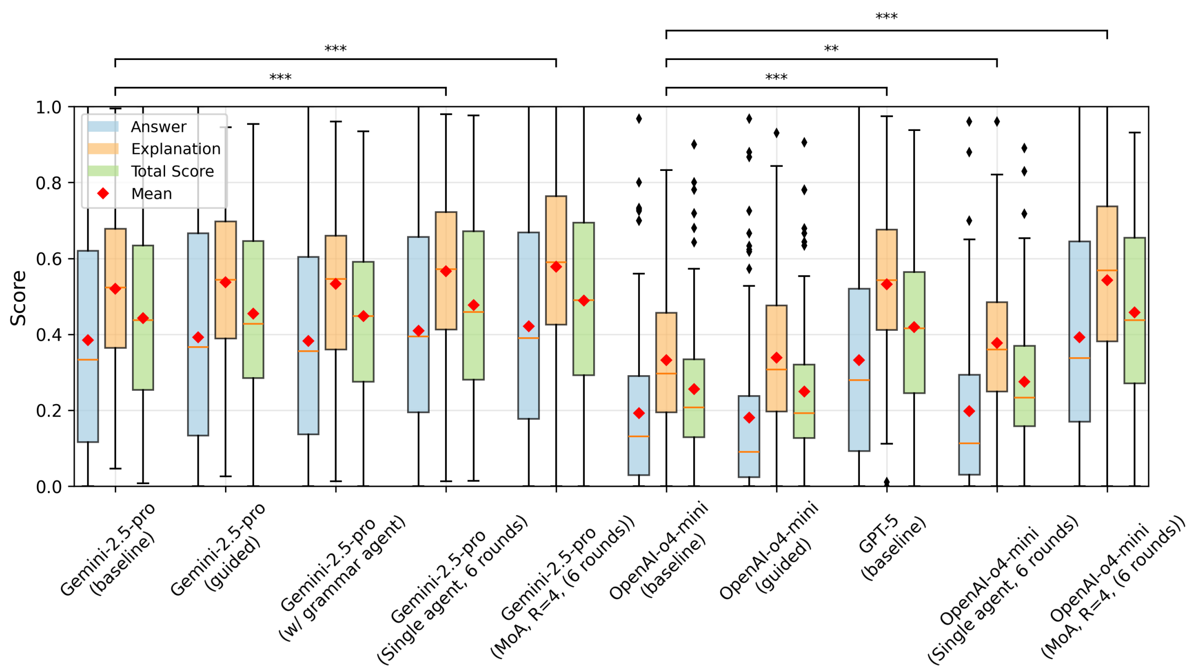


Figure 2: Scores on LOBSTER🦞of Different Models. Statistical significance was examined using paired Student's t-test. For simplicity, we only plot the significance between baseline vs. other models. {*,**,***} denotes $p <$ $\{0.05, 0.005, 0.0005\}$, respectively. The model name in MoA denotes the final aggregator and R is the number of intermediate rounds.

pro and OpenAI-o4-mini. The former considerably outperforms the latter, and is marginally better than GPT-5.

The trends between different settings are less clear: we found no statistically significant difference in the grammar agents scores compared to the baseline, nor in guided prompts vs. baseline. These results contradicts our expectation of an improvement; for discussions on possible reasons, see Section 5.3.

On the other hand, Mixture-of-Agents gives steadily increasing scores as the number of rounds increases, which are significantly better ($p < 0.05$) than the baseline as long as there is more than one round. Interestingly, the final aggregator plays an important role in the performance—if the final aggregator is weak (in this case, OpenAI-o4-mini), even though it has seen the (better) solutions generated by other models (in this case, Gemini-2.5-pro), the output scores far lower than the stronger model.

A natural question arises as to whether the effectiveness of MoA comes from multi-round from multi-agent. We introduced the single-agent multi-round setting to isolate their effects. Results show that additional rounds consistently improve performance, confirming the benefit of iterative reasoning. The multi-agent effect, however, is less pronounced for Gemini-2.5-pro—likely because it is already a stronger model, and a weaker collaborator offers limited help ($p = 0.105$ for 6-round MoA vs. single-agent multi-round with Gemini-2.5-pro). In contrast, OpenAI-o4-mini benefits greatly when paired with Gemini-2.5-pro ($p < 0.0001$).

The exact scores can be found in Appendix M.

5.2.2 Performance regarding Language Family and Problem Type

To gain a more nuanced breakdown of the model's performance, we analyzed the Gemini-2.5-Pro statistics by categorizing the problems based on language family, linguistics subject (e.g., phonology, syntax), and problem type (e.g., *Pattern*, *Match-up*). The detailed scores are plotted in the Appendix N.

Typologically, the model performs best on language isolates (mean = 0.70), Turkic (0.64), and Indo-European (0.55) languages, but struggles with Papuan (0.29), South American (0.25), and Australian (0.34) ones. The trend may be partially attributed to the resource-level of the languages.

By problem type, the model achieves its highest scores on *Monolingual* problems and lowest on *Match-up*. Across linguistic domains, it performs worst on syntax and best on semantics. The "Others" category has a score surpassing all others, possibly due to intrinsic differences in problem design.

Overall, the model shows strong performance in certain areas but inconsistent reasoning across languages, subjects, and problem types.

5.3 Discussions and Limitations

Exposure to the target language during pretraining. Even though the languages are low-resourced, models may still have some prior exposure that gives them an advantage in problem solving, meaning scores may not reflect pure reasoning ability. Additionally, the Internet presence of IOL problems increased the possibility of being in the pretraining data for some models. One approach to mitigate this is to systematically adjust the orthography, making it harder for models to recognize the language while preserving the problem's content (Khouja et al., 2025). Our work provides a solid foundation well-suited for future use.

Unimodality. Currently, the benchmark is designed to handle only text, in order to be applicable to a wider range of models. However, linguistics problems may involve other modalities (e.g., visual data), as seen in problems involving writing systems, kinship trees, and even maps. Such problems could be transcribed into text if possible but are usually excluded from the benchmark.

The exact content of the Grammar Agent. Contrary to our expectation, we found no major improvement when a model was equipped with a reference book. Dissecting the reason for this observation is a non-trivial task because the content and format of reference grammar books vary greatly, creating many confounding variables. For example, as Aycock et al. (2025) have discovered, the example sentences may be more useful than long paragraphs of grammar descriptions.

Another possible reason lies in the complexity of language itself. Reference grammar books are not a unified or accurate reflection of language but rather artifacts that attempt to summarize the real-world language use. Consequently, for the same language, it is not uncommon for different sources to have different orthographical conventions for transcription, variations from the data (e.g., speaker/dialect variations), and conflicting theories about grammar, where later works may disagree with the past literature. In Tanzer et al. (2024), these inconsistencies did not emerge, and we hypothesize that this is because their work used the same, consistent source for benchmarking and knowledge provision.

In any case, investigating the nature of external knowledge is necessary to continue the study. Such studies may require high-quality classification and annotation of books broken down into meaningful units, which we anticipate will demand considerable manual effort.

Reasoning traces. While our benchmark is a leap forward from previous linguistic reasoning benchmarks (in particular, ours is able to evaluate partially correct solutions meticulously, and is rich in metadata), the "thought process" of a model is not taken into consideration when grading. To our knowledge, evaluating the reasoning steps of LLMs remains an open problem.

To help advance this line of research, we provide a dataset of the gold-standard reasoning traces alongside the quantitative grading part of the benchmark, and ensure that their formats are fully compatible. One possible quantitative use of the reasoning trace data is as a "rule checklist," similar to the explanation grading in Section 4.3. This dataset, for which direct applications are yet to be explored, invites future researchers interested in reasoning and human cognition.

6 Conclusion

In this work, we introduced LOBSTER🦞, a linguistically-informed benchmark designed to move beyond final-answer accuracy and enable a granular assessment of an LLM's reasoning on complex linguistic structures. Our typological analysis of IOL problems provides a structured lens for this evaluation, while our empirical study of a state-of-the-art model on the FLORES-200 dataset underscored the critical need for improved cross-linguistic generalization, particularly in low-resource settings. We call on the community to build on this foundation to look inward at the nascent logic of LLMs, and outward at the boundless diversity of language that inspires them.

Acknowledgments

We would like to thank Zi-Cheng Yang for his help with implementation, Tung-Le Pan (Chair of Jury of IOL 2025) for comments about the IOL, and the members of Taiwan Olympiad in Linguistics—especially Hung-Chi Chen, Yin-Shuo Chang, Kanoa Ziyang Teng and Chloe Cheng—for the annotation and refinement of the data. We would also like to thank the reviewers for their valuable feedback.

References

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can LLMs really learn to translate a low-resource language from one grammar book? In *The Thirteenth International Conference on Learning Representations*.

Andrew Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A., Ryan Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 26224–26237. Curran Associates, Inc.

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 1–8, Sofia, Bulgaria. Association for Computational Linguistics.

Nathan Chi, Teodor Malchev, Riley Kong, Ryan Chi, Lucas Huang, Ethan Chi, R. McCoy, and Dragomir Radev. 2024. ModeLing: A novel dataset for testing linguistic reasoning in language models. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 113–119, St. Julian's, Malta. Association for Computational Linguistics.

Stanton P. Durham and David Ellis Rogers. 1969. An application of computer programming to the reconstruction of a proto-language. In *International Conference on Computational Linguistics COLING 1969: Preprint No. 5*, Sånga Säby, Sweden.

Kuofeng Gao, Huanqia Cai, Qingyao Shuai, Dihong Gong, and Zhifeng Li. 2025. Embedding self-correction as an inherent ability in large language models for enhanced mathematical reasoning.

Satyam Goyal and Soham Dan. 2025. Iolbench: Benchmarking llms on linguistic reasoning.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.1. *Leipzig: Max Planck Institute for Evolutionary Anthropology.(Available online at glottolog.org, Accessed on 2025-02-06.)*, 10.

Jiayi He, Hehai Lin, Qingyun Wang, Yi Fung, and Heng Ji. 2025. Self-correction is more than refinement: A learning framework for visual and language reasoning tasks.

IOL. 2003-2025. International linguistics olympiad.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, Caiming Xiong, and Shafiq Joty. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems.

Jude Khouja, Karolina Korgul, Simi Hellsten, Lingyi Yang, Vlad Neacsu, Harry Mayne, Ryan Kearns, Andrew Bean, and Adam Mahdi. 2025. Lingoly-too: Disentangling reasoning from knowledge with templatised orthographic obfuscation.

Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. 2025. Rethinking mixture-of-agents: Is mixing different large language models beneficial?

Liang Lu, Peirong Xie, and David Mortensen. 2024. Semisupervised neural proto-language reconstruction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14715–14759, Bangkok, Thailand. Association for Computational Linguistics.

Atharva Naik, Kexun Zhang, Nathaniel Robinson, Aravind Mysore, Clayton Marr, Hong Sng Rebecca Byrnes, Anna Cai, Kalvin Chang, and David Mortensen. 2024. Can large language models code like a linguist?: A case study in low resource sound law induction.

Vlad A. Neacşu. 2024. *Linguistics Olympiad*. Number 13 in Textbooks in Language Sciences. Language Science Press, Berlin.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. 2023. Art: Automatic multi-step reasoning and tool-use for large language models.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online. Association for Computational Linguistics.

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.

Yu-Fei Shih, Zheng-Lin Lin, and Shu-Kai Hsieh. 2025. Reasoning over the glyphs: Evaluation of llm's decipherment of rare scripts.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Linguini: A benchmark for language-agnostic linguistic reasoning.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *International Conference on Representation Learning*, volume 2024, pages 18955–18985.

Saujas Vaduguru, Aalok Sathe, Monojit Choudhury, and Dipti Sharma. 2021. Sample-efficient linguistic generalizations through program synthesis: Experiments with phonology problems. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 60–71, Online. Association for Computational Linguistics.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. 2025. Mixture-of-agents enhances large language model capabilities. In *International Conference on Representation Learning*, volume 2025, pages 33944–33963.

Mengsong Wu, Tong Zhu, Han Han, Xiang Zhang, Wenbiao Shao, and Wenliang Chen. 2025. Chain-of-tools: Utilizing massive unseen tools in the cot reasoning of frozen language models.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

A IOL Problem Examples

A.1 Elements of an IOL Problem

Problem 1 (20 points). Here are some forms of the Ubykh verb *to give* and their English translations:

Introduction

1. ***wəš̓’tʷən*** — *we give you$_{sg}$ to him*
2. ***sawtʷən*** — *you$_{sg}$ give me to them*
3. ***awəstʷan*** — *I give them to you$_{sg}$*
4. ***wasənatʷən*** — *they give you$_{sg}$ to me*
5. ***ŝʷəstʷan*** — *I give you$_{pl}$ to him*
6. ***š̓’antʷan*** — *he gives us to them*
7. ***awəš̓’tʷən*** — *we give him to you$_{sg}$*
8. ***səŝʷəntʷan*** — *he gives me to you$_{pl}$*
9. ***aŝʷəstʷan*** — *I give him to you$_{pl}$*

Corpus

(a) The last of the nine forms above can actually be translated into English in two ways. What is its other translation?

(b) Translate into English:

10. ***aš̓’əntʷən***
11. ***səŝʷtʷan***
12. ***š̓’əwənatʷan***

Tasks

(c) Translate into Ubykh:

13. *they give you$_{pl}$ to me*
14. *you$_{pl}$ give him to me*
15. *you$_{sg}$ give us to him*
16. *we give you$_{sg}$ to them*
17. *he gives them to us*

Notes

⚠ Ubykh belongs to the Abkhaz–Adyghe family. Until 1864, several tens of thousands of people spoke it in the area of the present-day city of Sochi, Russia. Tevfik Esenç, who was considered the last fully proficient native speaker of Ubykh, died in Turkey in 1992.

ə is a vowel; ***š̓’***, ***ŝʷ***, ***tʷ*** are consonants. —*Peter Arkadiev*

Figure 3: An IOL Problem with the four parts: **Introduction** provides information about the language(s) featured in the problem; **Corpus** contains the examples based on which the tasks should be solved. **Tasks** follows the corpus and typically includes translation between the target language and English, correspondences of randomly arranged items, among other types of tasks; **Notes** provide data about the language featured in the problem, relevant phonetic/orthographic information, and details about specific words. Any additional information crucial to solving the problem will be included in the **introduction** and **notes** sections.

+

A.2 More Examples on Diversity in Problem

Problem 1 (20 points). Here are some arithmetic equalities in Birom:

1. **tùŋūn**^{2} + **tàt** + **nààs** = **bākūrū bībā ná vὲ rwīīt**
2. **tàt**$^{\textbf{nààs}}$ = **bākūrū bītīīmìn ná vὲ ʃāātàt**
3. **tàāmà**^{2} + **ʃāātàt** + **gwīnìŋ** = **bākūrū bīnāās ná vὲ ʃāāgwīnìŋ**
4. **ʃāātàt**$^{\textbf{gwīnìŋ}}$ = **ʃāātàt**
5. **rwīīt**^{2} + **bà** + **tùŋūn** = **bākūrū bītūŋūn ná vὲ ʃāāgwīnìŋ**
6. **bà**$^{\textbf{tùŋūn}}$ = **bākūrū bībā ná vὲ rwīīt**
7. **ʃāātàt**^{2} + **nààs** + **tàt** = **bākūrū bītāāmà ná vὲ nààs**
8. **nààs**$^{\textbf{tàt}}$ = **bākūrū bītūŋūn ná vὲ nààs**
9. **kūrū ná vὲ nààs** + **kūrū ná vὲ ʃāātàt** = **kūrū ná vὲ tììmìn** + **bà** + **kūrū ná vὲ tùŋūn**

All numbers in this problem are greater than 0 and less than 125.

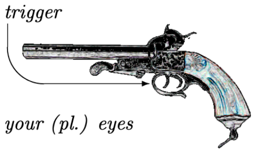
(a) Write the equalities (1–9) in numerals.

Figure 4: Problem 1 (IOL 2017)

Problem 2 (20 points). Here are some words and word combinations in Abui and their English translations in arbitrary order:

1. **abang**
2. **atáng heya**
3. **bataa hawata**
4. **dekafi**
5. **ebataa hatáng**
6. **ekuda hawata**
7. **falepak hawei**
8. **hatáng hamin**
9. **helui**
10. **maama hefalepak**
11. **napong**
12. **rièng**
13. **ritama**
14. **riya hatáng**
15. **tama habang**
16. **tamin**
17. **tefe hawei**

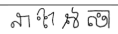
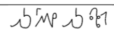
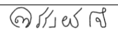
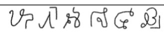
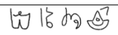
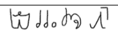
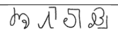
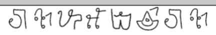
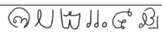
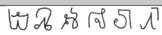
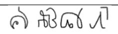
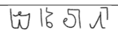
a. *his fingertip*
b. *your (sg.) branch*
c. *my face*
d. *one's own rope*
e. *your (sg.) shoulder*
f. *your (pl.) mother's hand*
g. *our pigs' ears*
(the ear of the pig of each of us)
h. *father's pistol*
i. *your (sg.) horse's neck*
j. *trigger*

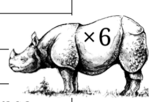


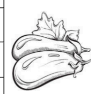
k. *your (pl.) eyes*
l. *our noses*
(the nose of each of us)
m. *his knife*
n. *seashore*
o. *upper part of a tree*
p. *your (sg.) thumb*
q. *your (pl.) sea*

Figure 5: Problem 2 (IOL 2017)

Problem 4 (20 points). Here are some word combinations in Laven written in the Khom script and in phonetic transcription and their English translations:

1		**praj trie**	*to wake up the wife*
2		**ca:k caj**	*from the heart/mind/soul*
3	?	**taw bɛː**	*to see the raft*
4		**krɨət blaw**	*to scratch the thigh*
5		**plaj prɨət**	*banana*
6	?	?	*three bananas*
7		?	*six rhinoceros*
8		?	*four hands of bananas*
9		?	*?*
10	?	**cie pʌh laː**	*seven sheets of paper*
11		?	*aubergine/eggplant leaf*
12		?	*two aubergines/eggplants*
13		**plaj hnat pʌh plaj**	*seven pineapples*
14		**kruat pɛː toː**	*three bees*
15		**laː prɨət traw laː**	*?*
16	?	**kəːr bəːr toː**	*two doves*
17		**blaːk puan kaː**	*four carp*
18		**piet traw plaː**	*six knives*
19		**bəːr kaː**	*?*
20		?	*four blades*





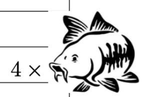
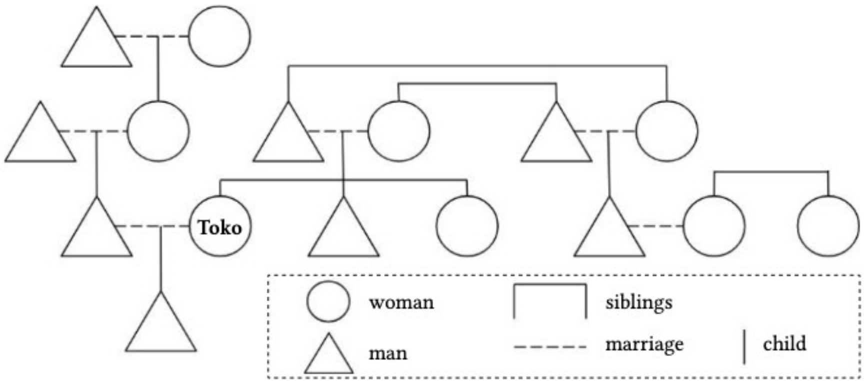


Figure 6: Problem 4 (IOL 2017)

A.3 Example on Kinship Problem&Graph Transcription

Twenty-first International Linguistics Olympiad (2024) 4
Individual Contest Problems

Problem 3 (20 points). You are given the family tree of a Komnzo-speaking family and statements describing the family members' relation to each other. Siblings are displayed in descending age order from left to right. The position of one family member, **Toko**, is known.



1. **Wafine Kuraiane nge rä.**
2. **Mea Gwamane bäiŋaf yé.**
3. **Naimr Tokoane ŋame rä.**
4. **Mea Wimsane ŋafe yé.**
5. **Marua Kuraiane enat yé.**
6. **Naimr Gwamane** ...①.
7. **Abia Maragaane ŋäwi yé.**
8. **Tawth Kuraiane zath yé.**
9. **Trafe Wafineane ŋame rä.**
10. **Marua Maragaane zath yé.**
11. **Tawth Meaane** ...②.
12. **Abia Gwamane yamit yé.**
13. **Tawth Wafineane nge yé.**
14. **Wafine Maragaane zath ŋare rä.**
15. **Kurai Wafineane ŋafe yé.**
16. **Trafe Tawthane** ...③.
17. **Mea Maragaane zath yé.**
18. **Nfiyam Wimsane bäiŋam rä.**
19. **Wims Gwamane yamit rä.**
20. **Maraga Tawthane** ...④.
21. **Skri Gwamane ŋafe yé.**
22. **Naimr Maragaane zath ŋare rä.**
23. **Maraga Tokoane nge yé.**
24. **Abia Tokoane ngth yé.**
25. **Toko Wimsane nane rä.**
26. **Toko Gwamane yamit rä.**
27. **Maraga Wafineane zath yé.**
28. **Nakre Wimsane yumad rä.**
29. **Abia Wimsane nane yé.**
30. **Mabata** ...⑤ **ngth** ...⑥.

(a) Fill in the family tree.

(b) Fill in the gaps (1–6).

(c) The following statement is incorrect. Explain why and correct the mistake.

31. **Skri Abiaane ŋäwi yé.**

⚠ The Komnzo language belongs to the Yam family. It is spoken by approx. 250 people in Rouku village and the town of Morehead in the Western Province of Papua New Guinea. The Farem people – the primary speakers of Komnzo – practice sister exchange, whereby two men from different clans marry each other's sisters (as seen in the family tree).

ä = *a* in *cat*. **ŋ** = *ng* in *hang*. **th** = *th* in *leather*. **z** = *ts* in *cats*. —*Aida Davletova*

Figure 7: Original Problem 3 in 2024.

Transcription of the Family Tree

- Man 1 and Woman 1 are married. Their child is Woman 2.
- Man 2 and Woman 2 are married. Their child is Man 3.
- Man 3 and Woman 3 are married. Their child is Man 4.
- Woman 3 is Toko.
- Man 5 and Woman 4 are married. Their children are Woman 3, Man 6 and Woman 5, from oldest to youngest.
- Man 5 and Woman 6 are siblings. The former is older.
- Woman 4 and Man 7 are siblings. The former is older.
- Man 7 and Woman 6 are married. Their child is Man 8.
- Man 8 and Woman 7 are married.
- Woman 7 and Woman 8 are siblings. The former is older.

B Prompt Template for Reasoning Process Generation

The following Python template was used to generate reasoning chains for IOL problems:

```
## Prompt:
As an expert in linguistics solve the following problem. Given the following IOL problem and its answer, generate a detailed, step-by-step chain of thoughts that could specifically and reasonably lead to the answer. Focus on the reasoning process, essential linguistic rules, logical deductions, and the final solution. Make your whole output into a markdown file.

## Problem:
{problem_text}

## Solution:
{solution_text}

## Your response:
```

C The Classification Framework for Problems

Category	Tag
Subject	Compounding, Morphology, Numbers, Phonology and Phonetics, Semantics, Syntax, Writing System
Type	Rosetta, Match-up, Monolingual, Pattern, Computational, Text
Theme	Classical, Comparative, Encrypted, Kinship, Maps, Mystery, MFL[1] Senses and Feelings, Stories, Poetry, No Theme

[1] MFL: These questions involve languages commonly taught in secondary school MFL departments, or those closely related (e.g., Romance and Germanic languages).

Table 1: Typological Annotation Category

Sub-problems	Subject	Type	Language	Speakers	glottocode	Language Family
2	Numbers	Pattern	Egyptian Arabic	68,000,000	egyp1253	Semitic

Table 2: Example of Typological Annotation: Problem 2 in 2003

C.1 Classification Criteria

The following categories and the classification criteria are modified from those of UKLO[9].

- **Subjects** –For a given subject to appear in the classification, at least two rules in the solution must be of that type.
  - **Compounding**: The problems mainly focus on deducing the dictionary meanings of words by analyzing how the meaning changes when different word components are combined.
  - **Morphology**: The problems primarily require understanding how morphemes (the smallest units of meaning) combine to form grammatical words.
  - **Numbers**: The problems are centered on understanding the structure and formation of numerals and numeral expressions.
  - **Phonology and Phonetics**: The problems focus on the sounds of a language and how they are organized. Phonology deals with sound systems within specific languages and in general, while phonetics studies the nature, production, and perception of speech sounds, independent of any particular language.

[9] https://www.uklo.org/technical-information/

- **Semantics**: The problems emphasize understanding how meaning influences language, especially how meaning shapes grammar and how different languages express the same concepts with different words.
- **Syntax**: The problems focus on understanding how words combine to form phrases and sentences.
- **Writing System**: The problems involve analyzing writing systems, including both the use of the Latin alphabet in various languages and other scripts.

• **Problem Type**

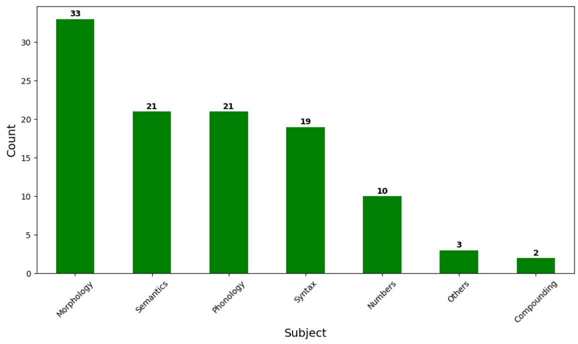
- **Rosetta**: The problems consist of sets of corresponding words or phrases across different languages or writing systems, with most pairings provided. Some elements may be missing, creating gaps that need to be filled. Solving the task requires generating new correspondences, typically translations.
- **Match-up**: The problems consist of sets of corresponding words or phrases across multiple languages or writing systems, with only a few pairings given. Some words may not belong to any set, but it still qualifies as a match-up. The task involves identifying new correspondences, usually translations.
- **Monolingual**: The problems are texts in an unfamiliar language (or equivalent writing system), generally without direct translations or transliterations, except perhaps for one or two words. To solve the task, you must translate the text from the unknown language.
- **Pattern**: The problems consist of words or groups of word forms or cognates that follow a certain pattern, though there may be exceptions. To solve the task, you must either generate other examples that fit the pattern or identify exceptions, without relying on translation as in Rosetta tasks.
- **Computational**: The problems include a description of a computational or logical system. Solving the problem involves analyzing and implementing the system according to the given rules.
- **Text**: The problems consist of full texts in different languages or scripts, without being broken down into smaller parts. To solve the task, you must infer linguistic rules using context and other cues.

• **Theme**

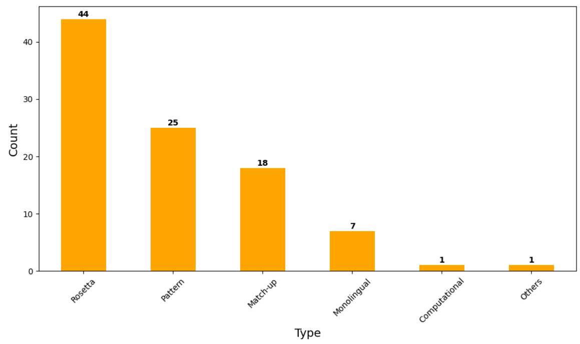
- **Classical**: These problems feature languages that were primarily spoken around a thousand years ago or earlier.
- **Comparative**:These problems involve comparing either related languages or different historical stages of a single language.
- **Encrypted**: These problems involve deciphering an encoded message in English.
- **Kinship**: These problems focus on understanding how different languages and cultures describe family relationships and naming systems.
- **Maps**: These problems explore how various languages express and conceptualize directions and spatial orientation.
- **Mystery**: These problems include a mystery element that draws on general or world knowledge, often involving content beyond linguistics.
- **MFL**: These problems involve languages commonly taught in secondary school modern foreign language (MFL) departments, or closely related languages (e.g., those from the Romance or Germanic families).
- **Senses and Feelings**: These problems examine linguistic expressions related to emotions or sensory experiences (e.g., smells, sounds).

– **Stories**: These problems either contain a narrative storyline or feature one or more fictional characters. They use storytelling to create engaging contexts for linguistic analysis, often drawing from literary traditions.
– **Poetry**: These problems revolve around the structure and features of poetic language.
– **No Theme (N/A)**: These problems focus on core linguistic topics without any specific thematic context.

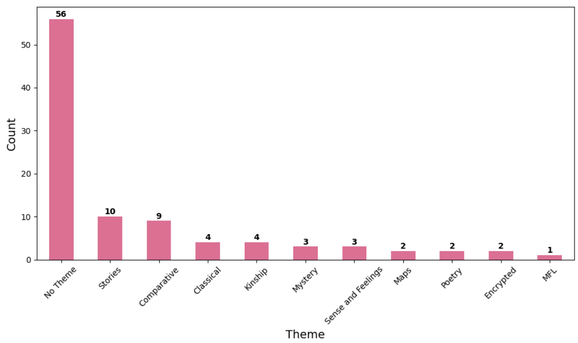
D Preliminary Analysis of IOL Problems.



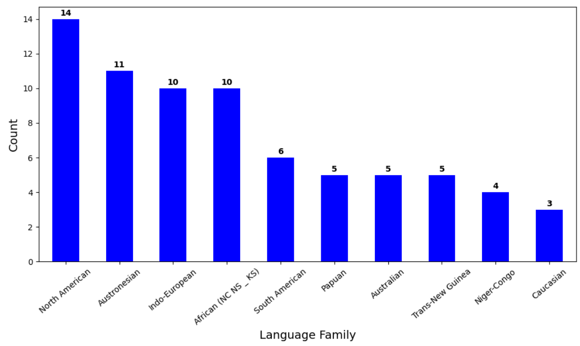
(a) Subject Distribution



(b) Type Distribution



(c) Theme Distribution



(d) Language Family Distribution

Figure 8: Statistical distributions of various features in the IOL problems dataset.

E Heatmap: Subject vs Type Distribution

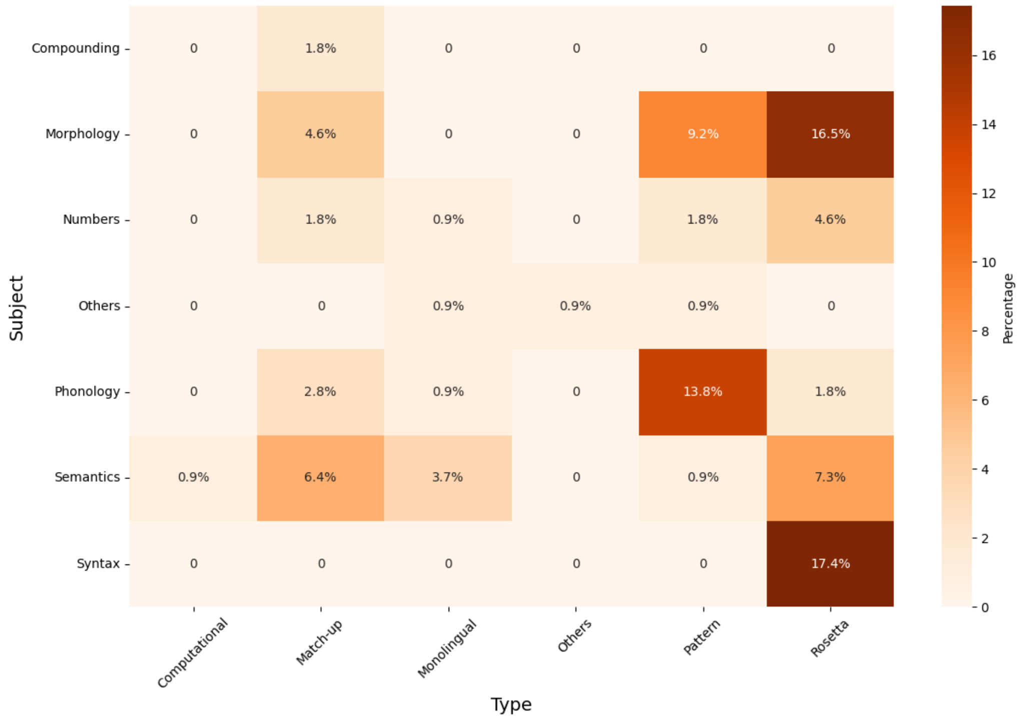


Figure 9: Subject vs Type Distribution

F Heatmap: Subject vs Language Family Distribution

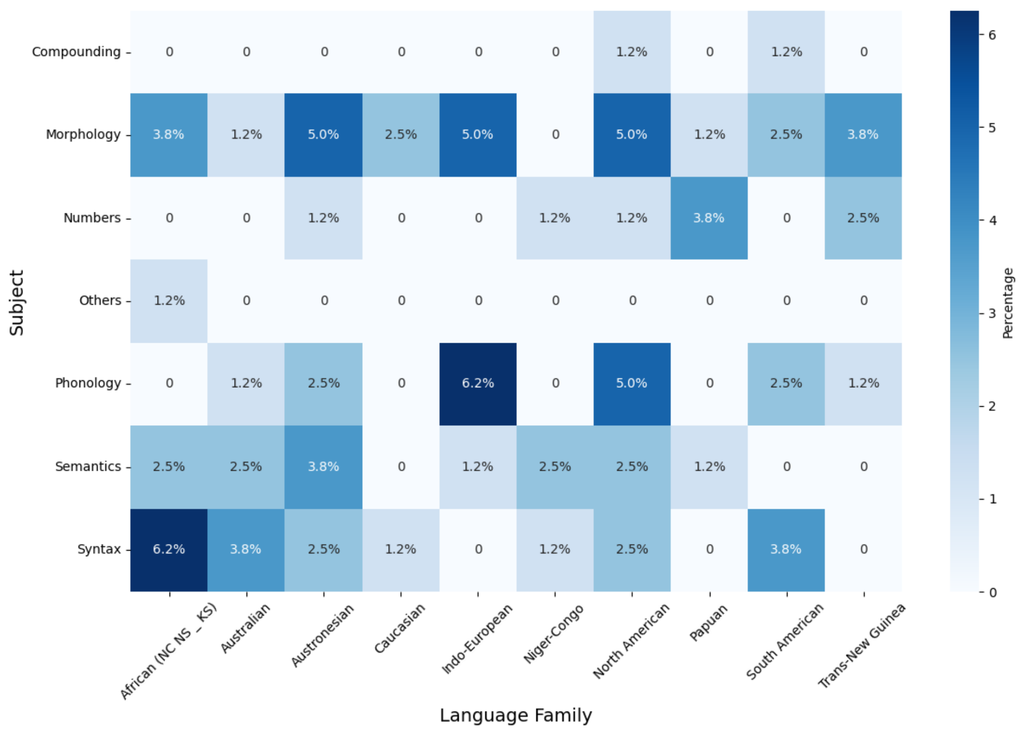


Figure 10: Subject vs Top 10 Language Family Distribution

G Heatmap: Type vs Language Family Distribution

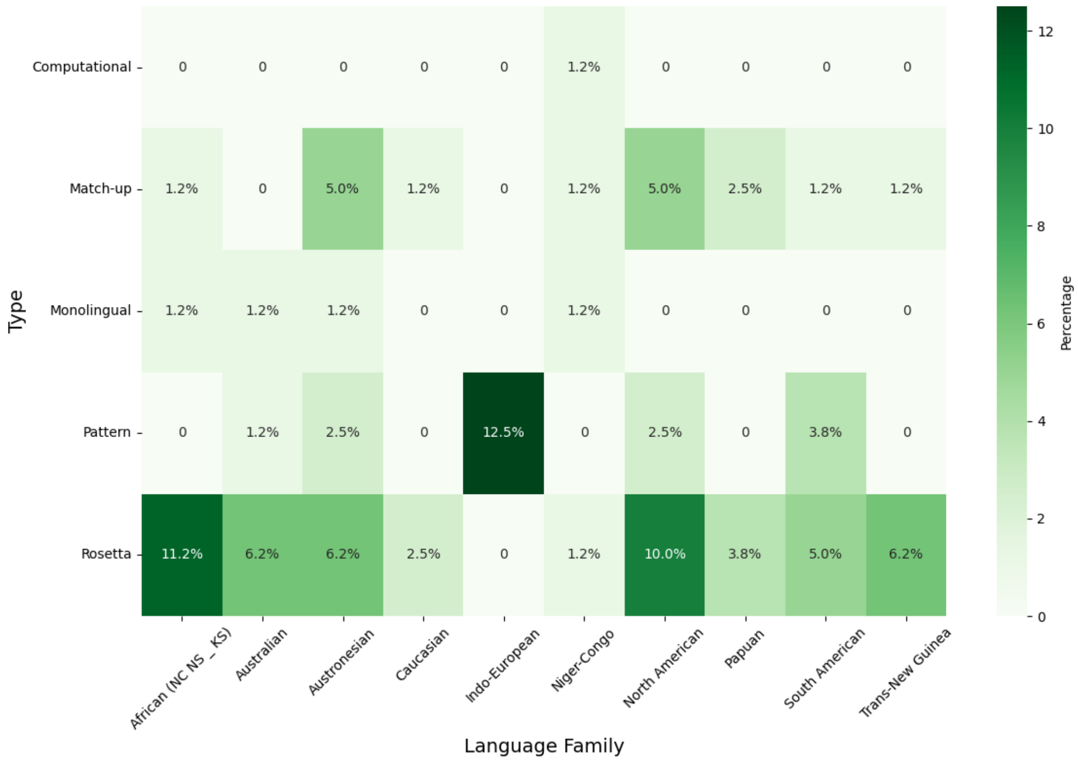


Figure 11: Type vs Top 10 Language Family Distribution

H System Prompt for Model Reasoning Evaluation

```
system_prompt = """Given the evaluation rules and metrics for model reasoning of
    IOL problems, consider the golden reasoning reference, and evaluate the target
    model reasoning with the metrics of five dimensions.
evaluation rules and metrics (5-score):
{metrics}

scoring_:
{scoring}

golden reasoning reference:
{golden_reasoning_reference}

target model reasoning:
{model_reasoning}
"""

metrics = """
### Metrics and Descriptions (Bullet Points)
 (i) 3.1 Information Extraction & Structuring
  * **Stepwise Logical Validity Score (SLVS)**: Measures whether each reasoning
    step is logically valid and aligned with the golden reasoning reference (GRR).
  * **Information Structuring Completeness (ISC)**: Measures completeness of
    extracted key information and its structure compared to GRR.
  ... [TRUNCATED FOR BREVITY IN PAPER]
"""

scoring = """
## Reasoning Quality Evaluation -- Scoring Rubric (5 Points per Metric)

| **Dimension** | **Metric** | **Score 5 (Excellent)** | **Score 3 (Acceptable)** |
     **Score 1 (Poor)** |
| ----- | ----- | ----- | ----- | ----- |
| **(i) 3.1 Information Extraction & Structuring** | **SLVS**   | All reasoning
    steps are logically valid and follow GRR structure | Minor logical flaws or
    omissions; generally coherent | Major logical errors, incoherent or illogical
    steps | |  | **ISC**    | Extracts and structures all key information as per GRR
     | Extracts partial or incomplete key information | Fails to extract/structure
    key information |
... [TRUNCATED FOR BREVITY IN PAPER]
"""

golden_reasoning_reference = """
# Your response:

# Chain of Thought: Solving the Swift News Linguistics Problem

... [TRUNCATED FOR BREVITY IN PAPER]
"""

target_model_reasoning = """
**Solving the Linguistic Puzzle**

... [TRUNCATED FOR BREVITY IN PAPER]
"""
```

I FLORES-200 Multilingual Evaluation Details

Dataset preparation and experimental design. We combine the `dev` and `devtest` splits for a total of 2009 sentences that are available in **204** languages. We then use the ISO 639-3 language code and the ISO 15924 script code to identify the Glottocode and the script used for each language, respectively. For example, the column name for Bashkir translations written in Cyrillic is `sentence_bak_Cyrl`. To align the dataset with the Glottolog taxonomy, we mapped all language identifiers to their corresponding Glottolog codes. We noted that five ISO 639-3 codes from the dataset (i.e., `srd`, `est`, `kon`, `zho`, `grn`) were not directly linked to a Glottolog entry. We identified suitable entries manually. How we mapped these languages can be found in Table 6. In total, we have 204 languages and script combinations.[10] Next, we take the first **10** English sentences and their translations for a total of **2030** English-to-Target Language pairs.

We evaluate Gemini-2.5-flash with `temperature=0.1` and `thinking budget=0` by translating from two directions: *English-to-Target* ($E \rightarrow T$) and *Target-to-English* ($T \rightarrow E$). We use the following $E \rightarrow T$ prompt when eliciting a response from the model:

```
Translate the following sentence from English to {target_lang} using
the {script} script:
Input: {input_sentence}
```

We use the following $T \rightarrow E$ prompt:

```
Translate the following sentence {target_lang} to English:
Input: {input_sentence}
```

Language	**Glottocode**	**Class**	**Missing ($E \rightarrow T$)**	**Missing ($T \rightarrow E$)**	**Total Missing**
Tamasheq	tama1365	0	7	1	8
Nuer	nuer1246	0	6	2	8
Kabiyé	kabi1261	0	7	0	7
Southwestern Dinka	sout2832	—	6	1	7
Central Kanuri	cent2050	0	4	2	6
Fon	fonn1241	0	5	0	5
Chokwe	chok1245	—	2	1	3
Umbundu	umbu1257	0	3	0	3
Kamba (Kenya)	kamb1297	0	2	0	2
Sango	sang1328	1	2	0	2
South-Central Koongo	koon1244	1	2	0	2
Kimbundu	kimb1241	0	2	0	2
Bambara	bamb1269	1	2	0	2
Dyula	dyul1238	0	2	0	2
Mossi	moss1236	0	4	0	4
Southern Jinghpaw	kach1280	0	4	0	4
Shan	shan1277	0	4	0	4
Acehnese	achi1257	1	1	0	1
Ewe	ewee1241	1	1	0	1
Dzongkha	dzon1239	1	1	0	1
Central Aymara	cent2142	—	1	0	1
Ayacucho Quechua	ayac1239	—	1	0	1
Luba-Lulua	luba1249	0	1	0	1
Kabyle	kaby1243	1	1	0	1
Guarani	east2555	1	1	0	1
Wolof	nucl1347	2	1	0	1
Grand Total			**73**	**7**	**80**

Table 3: Counts of missing LLM Outputs by language and direction. *Class* refers to the taxonomy introduced in Joshi et al. (2020) in which 0 indicates extremely limited resources and 5 indicates an abundance of resources. "–" means that the language was not found in the taxonomy.

[10] 196 unique languages while Acehnese, Minangkabau, Banjar, Central Kanuri, Tamasheq, Standard Arabic, Kashmiri, and Mandarin each have two scripts.

With the LLM translating in two directions, we obtain 3800 responses; however, 80 responses are empty with the majority of them originating from the $E \rightarrow T$ task. We will first examine these failures.

The LLM often fails to output any text for low resource languages. From the results in Table 3 we can see that the data strongly suggests that the model's failure to generate output is directly linked to data resource scarcity. The *Class* column refers to the taxonomy introduced in Joshi et al. (2020) where Class 0 languages have a dearth of resources while the Class 5 languages are at the opposite end of the spectrum.[11] The vast majority of missing outputs are concentrated in languages designated as Class 0 (e.g., Tamasheq, Nuer, Kabiyé), which represents the lowest-resource tier in our dataset. "–" means that the language was not found in the taxonomy.

Furthermore, the model fails far more frequently in the *English-to-Target* direction (73 instances) than in the Target-to-English direction (7 instances). This indicates that the primary challenge is not the model's ability to process or analyze the target languages (i.e., $T \rightarrow E$), but rather its capacity to reliably *generate* text in them (i.e., $E \rightarrow T$). This strongly suggests limited training data in the target language. This conclusion is reinforced by the performance on higher-resourced languages. We will now examine the overall translation quality of the outputs.

LLM performance is heavily influenced by translation direction, language family, and resource availability. We use chrF (Popović, 2015) instead of chrF+ or chrF++ (Popović, 2017) because the former is language independent and tokenization independent, which is needed when many languages found in FLORES-200 may not have a robust tokenizer or even have one readily available. chrF measures translation quality by calculating character-level n-gram overlap F-score between the machine translation and the human translation. The latter two introduces word unigram and bigram overlap into the equation. We use the implementation provided by Hugging Face with default parameters,[12] which adopts the implementation from sacreBLEU (Post, 2018)[13] but with a slightly different input format.

Direction	Mean chrF Score	Correlation with Class (ρ)
$E \rightarrow T$	43.92	0.598
$T \rightarrow E$	64.27	0.466

Table 4: Mean chrF scores and their Spearman's correlation (ρ) with resource class for each translation direction.

Worth noting is the direction where the model is worse on average ($E \rightarrow T$) is also the direction where performance is more strongly influenced by resource availability (higher correlation, $\rho = 0.598$). This suggests that while translating into English has a relatively high performance floor, the model's ability to generate text in other languages is both lower on average and more vulnerable to data scarcity. Figure 12 paints a similar picture in which lower resource classes predictably have worse performance compared to languages with more resources. We also see that translating from English to another language exacerbates the problem.

To also see how language family and script influence translation quality we used three separate one-way ANOVAs for each translation direction ($E \rightarrow T$ and $T \rightarrow E$). The results, summarized in Table 5, indicate that both **family** and **class** have a large and highly significant effect on performance in both directions (all $p < .001$). In contrast, **script** was not found to be a statistically significant predictor of chrF score in either analysis.

The analysis reveals an important asymmetry in the influence of resource class. While significant in both cases, **class** accounts for a larger portion of the variance in $E \rightarrow T$ scores ($\eta_p^2 = .412$) than in $T \rightarrow E$ scores ($\eta_p^2 = .381$).

[11] Because the language name to class list from Joshi et al. does not use an ISO 639-3 or Glottocode, we can only use the name to identify which language is paired with which Glottocode. We only assign classes for unambiguous language names. For example, while "khmer" is found in the language to class list, we do not join it with "Central Khmer." There are 30 languages without an assigned Resource Class.

[12] https://huggingface.co/spaces/evaluate-metric/chrf

[13] https://github.com/mjpost/sacreBLEU#chrf--chrf

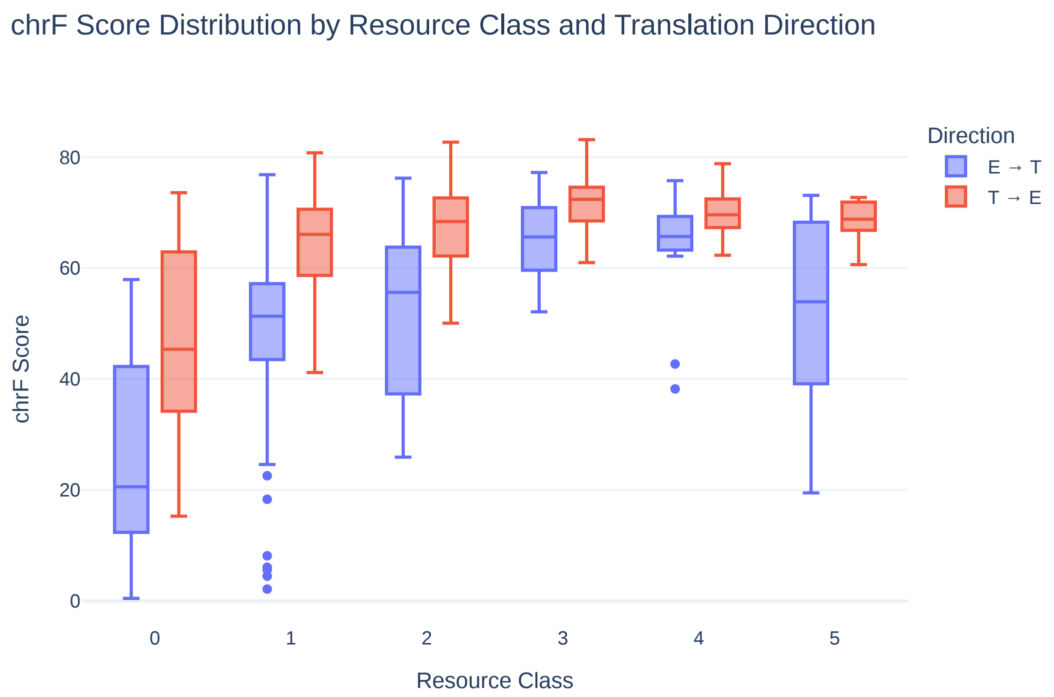


Figure 12: Comparison of CHRF score distributions for English-to-Target ($E \to T$) and Target-to-English ($T \to E$) translations, grouped by resource class. The plot shows a clear positive trend where quality increases with resource availability, with the $T \to E$ direction consistently outperforming the $E \to T$ direction. Boxes represent the interquartile range, and points show individual languages that fall beyond the lower fence.

	$E \to T$		$T \to E$	
Factor	**Effect Size (η_p^2)**	**p-value**	**Effect Size (η_p^2)**	**p-value**
Family	0.409	$< .001$	0.515	$< .001$
Class	0.412	$< .001$	0.381	$< .001$
Script	0.174	.265	0.125	.740

Table 5: Summary of One-Way ANOVA results showing the influence of each factor on CHRF scores. Effect sizes are given as partial eta-squared (η_p^2).

This illustrates that processing low-resource languages still proves to be a challenge for even the most powerful of models. FLORES-200 only covers a small fraction of the world's languages and were chosen carefully based on several considerations, such as having a presence on Wikipedia. This limitation with processing low-resource languages will only be more pronounced when we examine other languages with even fewer resources. The results for each language can be found in Table 7 as well as additional figures for script and language family-level scores in Section L of the Appendix.

Given that these results stem from a single experimental iteration, they should be interpreted as preliminary. Nevertheless, they provide strong evidence of the lopsided distribution of data resources among the world's languages and imbalanced performance across languages for today's SOTA LLMs, which warrants further investigation.

J Resolution of Ambiguous ISO 639-3 to Glottocode Mappings

Table 6: Resolution of ambiguous source ISO 639-3 codes to specific language varieties and their corresponding Glottocode.

Language Mapping Details	
`srd`	**Language:** Sardinian **ISO → Glottocode:** None → `sard1257` **Justification:** Top-level family node.
`est`	**Language:** Estonian **ISO → Glottocode:** `ekk` → `esto1258` **Justification:** Primary language entry.
`kon`	**Language:** South-Central Kongo **ISO → Glottocode:** `kng` → `koon1244` **Justification:** Known as Kongo in World Atlas of Language Structures (WALS).
`zho`	**Language:** Mandarin **ISO → Glottocode:** `cmn` → `mand1415` **Justification:** Most populous variety.
`grn`	**Language:** Eastern Bolivian Guaraní **ISO → Glottocode:** `gui` → `east2555` **Justification:** Guaraní categorized as Class 1 in Joshi et al. (2020), which aligns more with Ethnologue's Digital Language Support classification of "Ascending" for the language.

K Language-Level chrF Translation Scores for Gemini-2.5-Flash on FLORES-200

Table 7: Performance results by language, including chrF scores, sample counts, and resource class.

Language (glottocode_Script)	$E \rightarrow T$ chrF	$T \rightarrow E$ chrF	Family	Class	Samples ($E \rightarrow T$ / $T \rightarrow E$)
Acehnese (achi1257_Arabic)	6.05	49.46	Austronesian	1	9 / 10
Acehnese (achi1257_Latin)	46.42	71.11	Austronesian	1	10 / 10
Afrikaans (afri1274_Latin)	73.91	83.15	Indo-European	3	10 / 10
Akan (akan1250_Latin)	37.78	49.00	Atlantic-Congo	1	10 / 10
Amharic (amha1245_Ethiopic (Ge‘ez))	35.85	70.71	Afro-Asiatic	2	10 / 10
Assamese (assa1263_Bengali)	48.08	67.72	Indo-European	1	10 / 10
Asturian-Leonese-Cantabrian (astu1245_Latin)	69.93	73.94	Indo-European	1	10 / 10
Awadhi (awad1243_Devanagari (Nagari))	41.45	67.04	Indo-European	0	10 / 10
Ayacucho Quechua (ayac1239_Latin)	37.25	53.34	Quechuan	–	9 / 10
Balinese (bali1278_Latin)	44.79	61.53	Austronesian	0	10 / 10
Bambara (bamb1269_Latin)	2.12	41.72	Mande	1	8 / 10
Banjar (banj1239_Arabic)	4.46	53.69	Austronesian	1	10 / 10
Banjar (banj1239_Latin)	51.99	60.64	Austronesian	1	10 / 10
Bashkir (bash1264_Cyrillic)	56.01	68.67	Turkic	1	10 / 10
Basque (basq1248_Latin)	64.81	67.00	Unknown	4	10 / 10
Belarusian (bela1254_Cyrillic)	52.41	60.98	Indo-European	3	10 / 10
Bemba (Zambia) (bemb1257_Latin)	43.05	60.86	Atlantic-Congo	0	10 / 10
Bengali (beng1280_Bengali)	59.45	68.50	Indo-European	3	10 / 10
Bhojpuri (bhoj1244_Devanagari (Nagari))	44.14	62.46	Indo-European	1	10 / 10
Bosnian Standard (bosn1245_Latin)	67.72	70.89	Indo-European	3	10 / 10
Buginese (bugi1244_Latin)	35.98	48.74	Austronesian	1	10 / 10
Bulgarian (bulg1262_Cyrillic)	76.45	76.70	Indo-European	3	10 / 10
Burmese (nucl1310_Myanmar (Burmese))	53.93	68.35	Sino-Tibetan	1	10 / 10
Catalan (stan1289_Latin)	67.90	72.33	Indo-European	4	10 / 10
Cebuano (cebu1242_Latin)	65.84	80.13	Austronesian	3	10 / 10
Central Aymara (cent2142_Latin)	31.09	44.91	Aymaran	–	9 / 10
Central Kanuri (cent2050_Arabic)	2.31	15.26	Saharan	0	10 / 8
Central Kanuri (cent2050_Latin)	8.76	32.46	Saharan	0	6 / 10
Central Khmer (cent1989_Khmer)	43.45	73.44	Austroasiatic	–	10 / 10
Central Kurdish (cent1972_Arabic)	51.29	67.85	Indo-European	–	10 / 10
Central Moroccan Berber (cent2194_Tifinagh (Berber))	26.34	45.68	Afro-Asiatic	0	10 / 10
Chhattisgarhi (chha1249_Devanagari (Nagari))	50.58	70.19	Indo-European	–	10 / 10
Chokwe (chok1245_Latin)	19.24	28.74	Atlantic-Congo	–	8 / 9
Crimean Tatar (crim1257_Latin)	45.90	70.46	Turkic	1	10 / 10
Croatian Standard (croa1245_Latin)	62.14	69.88	Indo-European	4	10 / 10
Czech (czec1258_Latin)	63.42	73.96	Indo-European	4	10 / 10
Danish (dani1285_Latin)	77.23	75.40	Indo-European	3	10 / 10
Dari (dari1249_Arabic)	42.70	65.11	Indo-European	4	10 / 10
Dutch (dutc1256_Latin)	66.63	68.29	Indo-European	4	10 / 10
Dyula (dyul1238_Latin)	16.31	33.30	Mande	0	8 / 10
Dzongkha (dzon1239_Tibetan)	33.37	50.97	Sino-Tibetan	1	9 / 10
East Latvian (east2282_Latin)	45.29	73.02	Indo-European	–	10 / 10
Eastern Armenian (nucl1235_Armenian)	61.44	71.83	Indo-European	1	10 / 10
Eastern Panjabi (panj1256_Gurmukhi)	56.52	73.75	Indo-European	–	10 / 10
Eastern Yiddish (east2295_Hebrew)	42.70	83.12	Indo-European	–	10 / 10
Egyptian Arabic (egyp1253_Arabic)	52.11	65.85	Afro-Asiatic	3	10 / 10
Esperanto (espe1235_Latin)	66.90	76.29	Artificial Language	1	10 / 10
Estonian (esto1258_Latin)	61.18	67.24	Uralic	3	10 / 10
Ewe (ewee1241_Latin)	36.73	49.73	Atlantic-Congo	1	9 / 10
Faroese (faro1244_Latin)	64.14	77.47	Indo-European	1	10 / 10
Fijian (fiji1243_Latin)	50.32	55.60	Austronesian	1	10 / 10
Finnish (finn1318_Latin)	66.57	68.23	Uralic	4	10 / 10
Fon (fonn1241_Latin)	7.64	23.20	Atlantic-Congo	0	5 / 10
French (stan1290_Latin)	73.10	70.06	Indo-European	5	10 / 10
Friulian (friu1240_Latin)	61.78	67.92	Indo-European	1	10 / 10
Galician (gali1258_Latin)	65.03	70.14	Indo-European	3	10 / 10
Ganda (gand1255_Latin)	42.86	56.15	Atlantic-Congo	1	10 / 10
Georgian (nucl1302_Georgian (Mkhedruli))	56.55	63.11	Kartvelian	3	10 / 10
German (stan1295_Latin)	71.48	72.08	Indo-European	5	10 / 10

Continued on next page

Language (glottocode_Script)	$E \rightarrow T$ chrF	$T \rightarrow E$ chrF	Family	Class	Samples ($E \rightarrow T$ / $T \rightarrow E$)
Gilit Mesopotamian Arabic (meso1252_Arabic)	51.31	66.27	Afro-Asiatic	–	10 / 10
Guarani (east2555_Latin)	30.45	60.51	Tupian	1	9 / 10
Gujarati (guja1252_Gujarati)	49.30	70.17	Indo-European	1	10 / 10
Haitian (hait1244_Latin)	62.81	69.51	Indo-European	2	10 / 10
Halh Mongolian (halh1238_Cyrillic)	54.87	71.44	Mongolic-Khitan	0	10 / 10
Hausa (haus1257_Latin)	61.93	67.27	Afro-Asiatic	2	10 / 10
Hausa States Fulfulde (nige1253_Latin)	23.36	34.24	Atlantic-Congo	–	10 / 10
Hindi (hind1269_Devanagari (Nagari))	64.11	69.33	Indo-European	4	10 / 10
Hungarian (hung1274_Latin)	69.67	71.54	Uralic	4	10 / 10
Icelandic (icel1247_Latin)	65.36	69.58	Indo-European	2	10 / 10
Igbo (nucl1417_Latin)	50.62	64.71	Atlantic-Congo	1	10 / 10
Iloko (ilok1237_Latin)	56.05	69.03	Austronesian	1	10 / 10
Irish (iris1253_Latin)	64.73	77.31	Indo-European	2	10 / 10
Italian (ital1282_Latin)	62.85	64.59	Indo-European	4	10 / 10
Japanese (nucl1643_Japanese)	53.92	72.73	Japonic	5	10 / 10
Javanese (java1254_Latin)	64.70	71.11	Austronesian	1	10 / 10
Kabiyé (kabi1261_Latin)	0.44	39.03	Atlantic-Congo	0	3 / 10
Kabuverdianu (kabu1256_Latin)	58.01	75.68	Indo-European	–	10 / 10
Kabyle (kaby1243_Latin)	32.01	58.15	Afro-Asiatic	1	9 / 10
Kamba (Kenya) (kamb1297_Latin)	24.93	47.68	Atlantic-Congo	0	8 / 10
Kannada (nucl1305_Kannada)	55.88	63.90	Dravidian	1	10 / 10
Kashmiri (kash1277_Arabic)	26.62	62.82	Indo-European	1	10 / 10
Kashmiri (kash1277_Devanagari (Nagari))	22.55	57.73	Indo-European	1	10 / 10
Kazakh (kaza1248_Cyrillic)	64.98	72.01	Turkic	3	10 / 10
Kikuyu (kiku1240_Latin)	5.62	53.15	Atlantic-Congo	1	10 / 10
Kimbundu (kimb1241_Latin)	21.37	41.79	Atlantic-Congo	0	8 / 10
Kinshasa Lingala (ling1263_Latin)	48.36	53.12	Atlantic-Congo	1	10 / 10
Kinyarwanda (kiny1244_Latin)	59.06	65.75	Atlantic-Congo	1	10 / 10
Kirghiz (kirg1245_Cyrillic)	54.92	59.06	Turkic	1	10 / 10
Korean (kore1280_Hangul (Hangŭl, Hangeul))	38.21	62.31	Koreanic	4	10 / 10
Lao (laoo1244_Lao)	58.59	70.02	Tai-Kadai	2	10 / 10
Levantine Arabic (nort3139_Arabic)	67.19	74.01	Afro-Asiatic	–	10 / 10
Ligurian (ligu1248_Latin)	48.14	76.98	Indo-European	1	10 / 10
Limburgan (limb1263_Latin)	56.83	76.72	Indo-European	–	10 / 10
Lithuanian (lith1251_Latin)	65.99	71.04	Indo-European	3	10 / 10
Lombard (lomb1257_Latin)	40.32	67.99	Indo-European	1	10 / 10
Luba-Lulua (luba1249_Latin)	29.97	52.74	Atlantic-Congo	0	9 / 10
Luo (Kenya and Tanzania) (luok1236_Latin)	37.90	47.98	Nilotic	–	10 / 10
Macedonian (mace1250_Cyrillic)	64.95	70.12	Indo-European	1	10 / 10
Magahi (maga1260_Devanagari (Nagari))	57.93	73.59	Indo-European	0	10 / 10
Maithili (mait1250_Devanagari (Nagari))	50.43	66.99	Indo-European	1	10 / 10
Malayalam (mala1464_Malayalam)	59.07	69.10	Dravidian	1	10 / 10
Maltese (malt1254_Latin)	76.21	82.70	Afro-Asiatic	2	10 / 10
Mandarin (mand1415_Han (Simplified))	40.77	66.44	Sino-Tibetan	5	10 / 10
Mandarin (mand1415_Han (Traditional))	34.25	68.81	Sino-Tibetan	5	10 / 10
Manipuri (mani1292_Bengali)	19.06	64.31	Sino-Tibetan	0	10 / 10
Maori (maor1246_Latin)	47.45	64.97	Austronesian	1	10 / 10
Marathi (mara1378_Devanagari (Nagari))	52.66	66.06	Indo-European	2	10 / 10
Minangkabau (mina1268_Arabic)	8.12	61.44	Austronesian	1	10 / 10
Minangkabau (mina1268_Latin)	62.69	71.41	Austronesian	1	10 / 10
Mizo (lush1249_Latin)	50.39	59.40	Sino-Tibetan	0	10 / 10
Modern Greek (mode1248_Greek)	59.10	73.07	Indo-European	3	10 / 10
Modern Hebrew (hebr1245_Hebrew)	69.28	74.57	Afro-Asiatic	3	10 / 10
Moroccan Arabic (moro1292_Arabic)	45.14	60.62	Afro-Asiatic	5	10 / 10
Moselle Franconian (luxe1241_Latin)	59.83	75.58	Indo-European	1	10 / 10
Mossi (moss1236_Latin)	15.53	40.71	Atlantic-Congo	0	6 / 10
Najdi Arabic (najd1235_Arabic)	65.27	72.14	Afro-Asiatic	–	10 / 10
Nepali (nepa1254_Devanagari (Nagari))	52.28	70.34	Indo-European	1	10 / 10
North Azerbaijani (nort2697_Latin)	46.62	61.61	Turkic	–	10 / 10
Northern Kurdish (nort2641_Latin)	46.16	64.78	Indo-European	0	10 / 10
Northern Tosk Albanian (tosk1239_Latin)	64.32	74.24	Indo-European	–	10 / 10
Northern Uzbek (nort2690_Latin)	64.70	70.08	Turkic	–	10 / 10
Norwegian Bokmål (norw1259_Latin)	67.89	70.38	Indo-European	–	10 / 10
Norwegian Nynorsk (norw1262_Latin)	68.94	77.57	Indo-European	–	10 / 10

Continued on next page

Table 7 – continued from previous page

Language (glottocode_Script)	$E \rightarrow T$ chrF	$T \rightarrow E$ chrF	Family	Class	Samples ($E \rightarrow T$ / $T \rightarrow E$)
Nuer (nuer1246_Latin)	6.65	21.75	Nilotic	0	4 / 8
Nyanja (nyan1308_Latin)	57.28	64.36	Atlantic-Congo	1	10 / 10
Occitan (occi1239_Latin)	64.46	75.99	Indo-European	1	10 / 10
Odia (oriy1255_Oriya)	57.08	70.09	Indo-European	1	10 / 10
Pangasinan (pang1290_Latin)	50.29	67.22	Austronesian	1	10 / 10
Papiamento (papi1253_Latin)	59.40	77.99	Indo-European	1	10 / 10
Pedi (pedi1238_Latin)	58.90	72.17	Atlantic-Congo	–	10 / 10
Plateau Malagasy (plat1254_Latin)	54.33	66.30	Austronesian	1	10 / 10
Polish (poli1260_Latin)	63.24	68.08	Indo-European	4	10 / 10
Portuguese (port1283_Latin)	74.12	72.46	Indo-European	4	10 / 10
Romanian (roma1327_Latin)	72.24	73.24	Indo-European	3	10 / 10
Rundi (rund1242_Latin)	46.17	59.44	Atlantic-Congo	1	10 / 10
Russian (russ1263_Cyrillic)	70.87	69.89	Indo-European	4	10 / 10
Samoan (samo1305_Latin)	52.34	70.75	Austronesian	1	10 / 10
Sango (sang1328_Latin)	18.31	41.16	Atlantic-Congo	1	8 / 10
Sanskrit (sans1269_Devanagari (Nagari))	38.77	53.26	Indo-European	2	10 / 10
Santali (sant1410_Ol Chiki (Ol Cemet', Ol, Santali))	28.85	57.77	Austroasiatic	1	10 / 10
Sardinian (sard1257_Latin)	63.26	76.16	Indo-European	1	10 / 10
Scottish Gaelic (scot1245_Latin)	56.12	68.45	Indo-European	1	10 / 10
Serbian Standard (serb1264_Cyrillic)	63.79	74.85	Indo-European	4	10 / 10
Shan (shan1277_Myanmar (Burmese))	18.45	65.01	Tai-Kadai	0	6 / 10
Shona (shon1251_Latin)	50.02	53.16	Atlantic-Congo	1	10 / 10
Sicilian (sici1248_Latin)	50.63	68.78	Indo-European	1	10 / 10
Silesian (sile1253_Latin)	52.44	75.23	Indo-European	1	10 / 10
Sindhi (sind1272_Arabic)	56.57	71.45	Indo-European	1	10 / 10
Sinhala (sinh1246_Sinhala)	54.76	65.09	Indo-European	1	10 / 10
Slovak (slov1269_Latin)	59.60	68.26	Indo-European	3	10 / 10
Slovenian (slov1268_Latin)	70.90	72.76	Indo-European	3	10 / 10
Somali (soma1255_Latin)	48.80	62.48	Afro-Asiatic	1	10 / 10
South Azerbaijani (sout2697_Arabic)	37.49	63.69	Turkic	–	10 / 10
South Levantine Arabic (sout3123_Arabic)	53.99	70.58	Afro-Asiatic	–	10 / 10
South-Central Koongo (koon1244_Latin)	24.58	49.15	Atlantic-Congo	1	8 / 10
Southern Jinghpaw (kach1280_Latin)	21.18	45.03	Sino-Tibetan	0	6 / 10
Southern Pashto (sout2649_Arabic)	33.63	64.12	Indo-European	–	10 / 10
Southern Sotho (sout2807_Latin)	55.44	75.96	Atlantic-Congo	1	10 / 10
Southwestern Dinka (sout2832_Latin)	1.38	24.26	Nilotic	–	4 / 9
Spanish (stan1288_Latin)	63.33	66.93	Indo-European	5	10 / 10
Standard Arabic (stan1318_Arabic)	67.19	71.83	Afro-Asiatic	5	10 / 10
Standard Arabic (stan1318_Latin)	19.46	68.76	Afro-Asiatic	5	10 / 10
Standard Indonesian (indo1316_Latin)	74.66	69.53	Austronesian	3	10 / 10
Standard Latvian (stan1325_Latin)	63.66	73.36	Indo-European	3	10 / 10
Standard Malay (stan1306_Latin)	73.67	74.33	Austronesian	3	10 / 10
Sundanese (sund1252_Latin)	53.08	60.76	Austronesian	1	10 / 10
Swahili (swah1253_Latin)	75.19	77.87	Atlantic-Congo	2	10 / 10
Swati (swat1243_Latin)	47.46	59.26	Atlantic-Congo	1	10 / 10
Swedish (swed1254_Latin)	75.76	73.53	Indo-European	4	10 / 10
Ta'izzi-Adeni Arabic (taiz1242_Arabic)	57.90	68.61	Afro-Asiatic	–	10 / 10
Tagalog (taga1270_Latin)	65.38	79.03	Austronesian	3	10 / 10
Tajik (taji1245_Cyrillic)	57.78	65.02	Indo-European	1	10 / 10
Tamasheq (tama1365_Latin)	12.08	35.07	Afro-Asiatic	0	6 / 10
Tamasheq (tama1365_Tifinagh (Berber))	12.61	28.51	Afro-Asiatic	0	7 / 9
Tamil (tami1289_Tamil)	66.37	67.33	Dravidian	3	10 / 10
Tatar (tata1255_Cyrillic)	63.39	65.85	Turkic	1	10 / 10
Telugu (telu1262_Telugu)	58.70	74.29	Dravidian	1	10 / 10
Thai (thai1261_Thai)	64.09	74.75	Tai-Kadai	3	10 / 10
Tibetan (tibe1272_Tibetan)	46.95	58.32	Sino-Tibetan	1	10 / 10
Tigrinya (tigr1271_Ethiopic (Ge'ez))	26.43	61.36	Afro-Asiatic	2	10 / 10
Tok Pisin (tokp1240_Latin)	46.00	58.89	Indo-European	1	10 / 10
Tsonga (tson1249_Latin)	53.82	66.83	Atlantic-Congo	1	10 / 10
Tswana (tswa1253_Latin)	45.34	62.99	Atlantic-Congo	2	10 / 10
Tumbuka (tumb1250_Latin)	48.32	58.45	Atlantic-Congo	1	10 / 10
Tunisian Arabic (tuni1259_Arabic)	43.91	67.20	Afro-Asiatic	–	10 / 10
Turkish (nucl1301_Latin)	69.30	78.82	Turkic	4	10 / 10
Turkmen (turk1304_Latin)	54.86	67.57	Turkic	1	10 / 10

Continued on next page

Table 7 – continued from previous page

Language (glottocode_Script)	$E \rightarrow T$ chrF	$T \rightarrow E$ chrF	Family	Class	Samples ($E \rightarrow T$ / $T \rightarrow E$)
Twi (twii1234_Latin)	40.08	54.68	Atlantic-Congo	1	10 / 10
Uighur (uigh1240_Arabic)	57.10	63.85	Turkic	1	10 / 10
Ukrainian (ukra1253_Cyrillic)	67.63	73.64	Indo-European	3	10 / 10
Umbundu (umbu1257_Latin)	19.95	44.89	Atlantic-Congo	0	7 / 10
Urdu (urdu1245_Arabic)	56.80	69.39	Indo-European	3	10 / 10
Venetian (vene1258_Latin)	53.60	72.88	Indo-European	1	10 / 10
Vietnamese (viet1252_Latin)	68.50	67.29	Austroasiatic	4	10 / 10
Waray (Philippines) (wara1300_Latin)	61.97	80.62	Austronesian	1	10 / 10
Welsh (wels1247_Latin)	76.84	80.79	Indo-European	1	10 / 10
West Central Oromo (west2721_Latin)	43.92	58.33	Afro-Asiatic	–	10 / 10
Western Farsi (west2369_Arabic)	51.22	69.55	Indo-European	–	10 / 10
Wolof (nucl1347_Latin)	27.23	52.05	Atlantic-Congo	2	9 / 10
Xhosa (xhos1239_Latin)	51.60	64.15	Atlantic-Congo	2	10 / 10
Yoruba (yoru1245_Latin)	25.90	50.06	Atlantic-Congo	2	10 / 10
Yue Chinese (yuec1235_Han (Traditional))	30.09	68.45	Sino-Tibetan	1	10 / 10
Zulu (zulu1248_Latin)	58.61	74.58	Atlantic-Congo	2	10 / 10

L Supplementary Figures

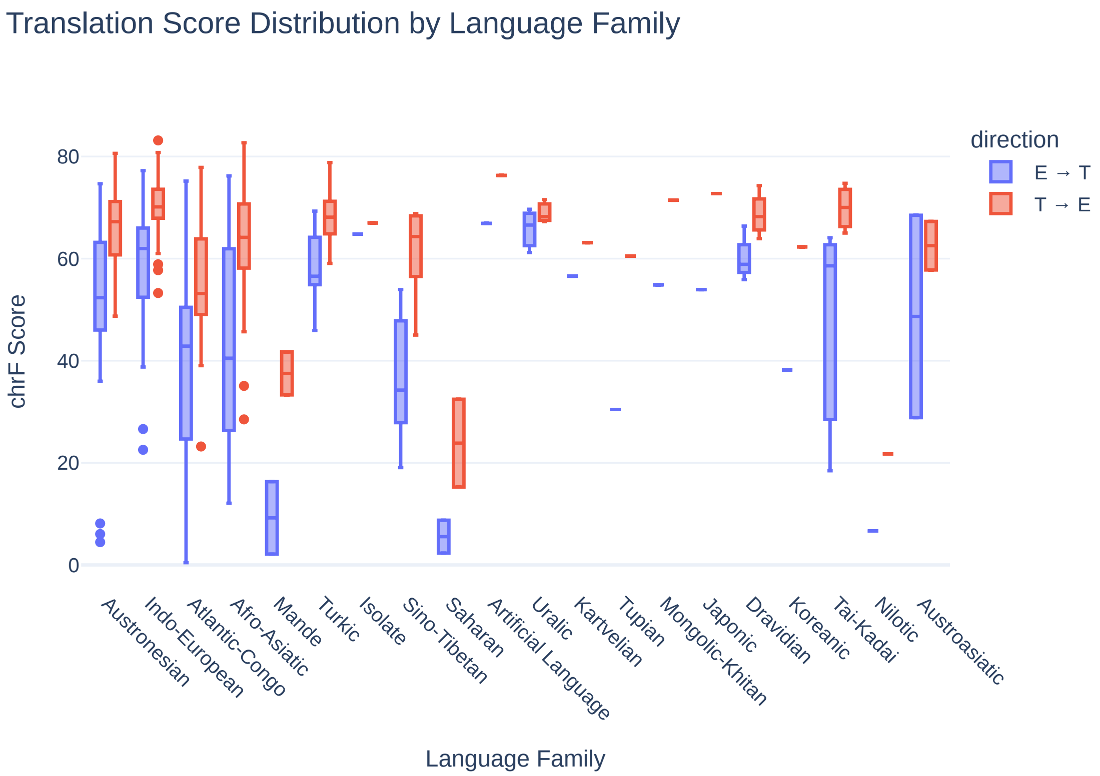


Figure 13: **Translation Score Distribution by Language Family.** This plot compares the distribution of chrF scores for English-to-Target ($E \rightarrow T$) and Target-to-English ($T \rightarrow E$) directions across language families. A consistent performance gap is evident, with $T \rightarrow E$ scores being almost universally higher and often less variable than $E \rightarrow T$ scores. Families such as Saharan and Mande show particularly low performance in the $E \rightarrow T$ direction, whereas families like Indo-European show a wider range of performance with generally higher scores.

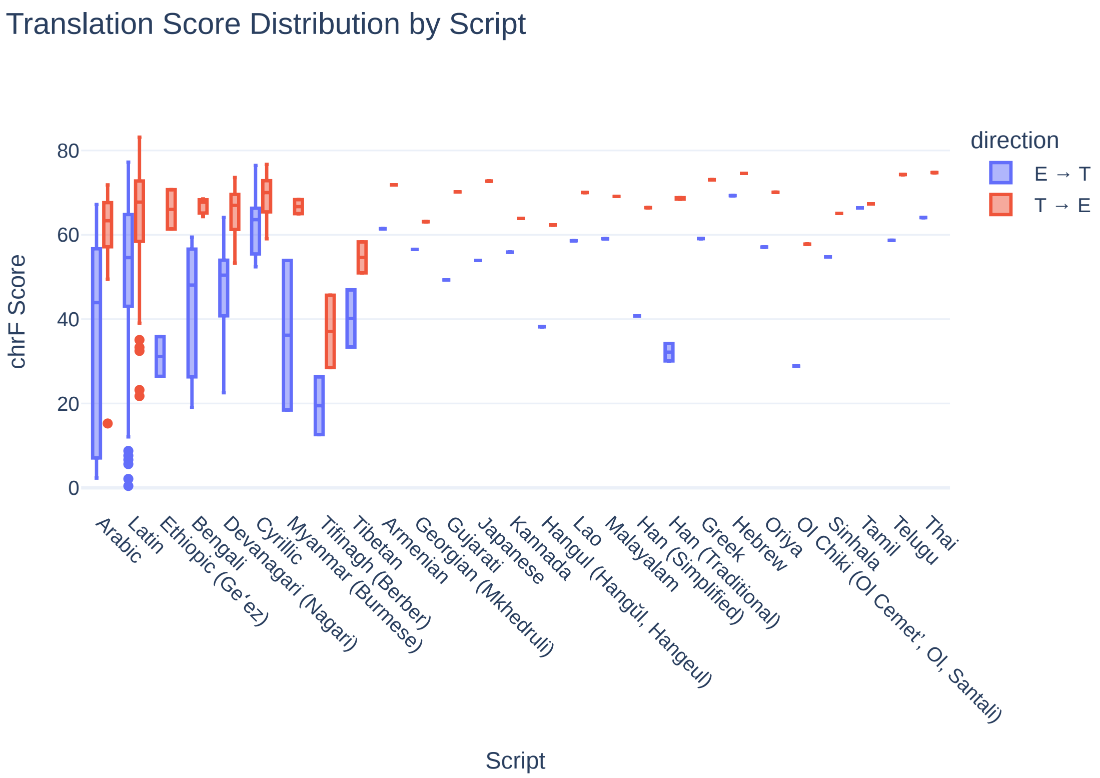


Figure 14: **Translation Score Distribution by Script.** This plot compares chrF score distributions across different writing systems. As with the family-based plot, the $T \rightarrow E$ direction consistently outperforms the $E \rightarrow T$ direction. Performance for languages using Latin and Cyrillic scripts is relatively high but shows a wide distribution, reflecting the diverse range of languages using them. Scripts associated with lower-resource languages, such as Ethiopic and Tifinagh, exhibit lower median scores, particularly in the $E \rightarrow T$ direction.

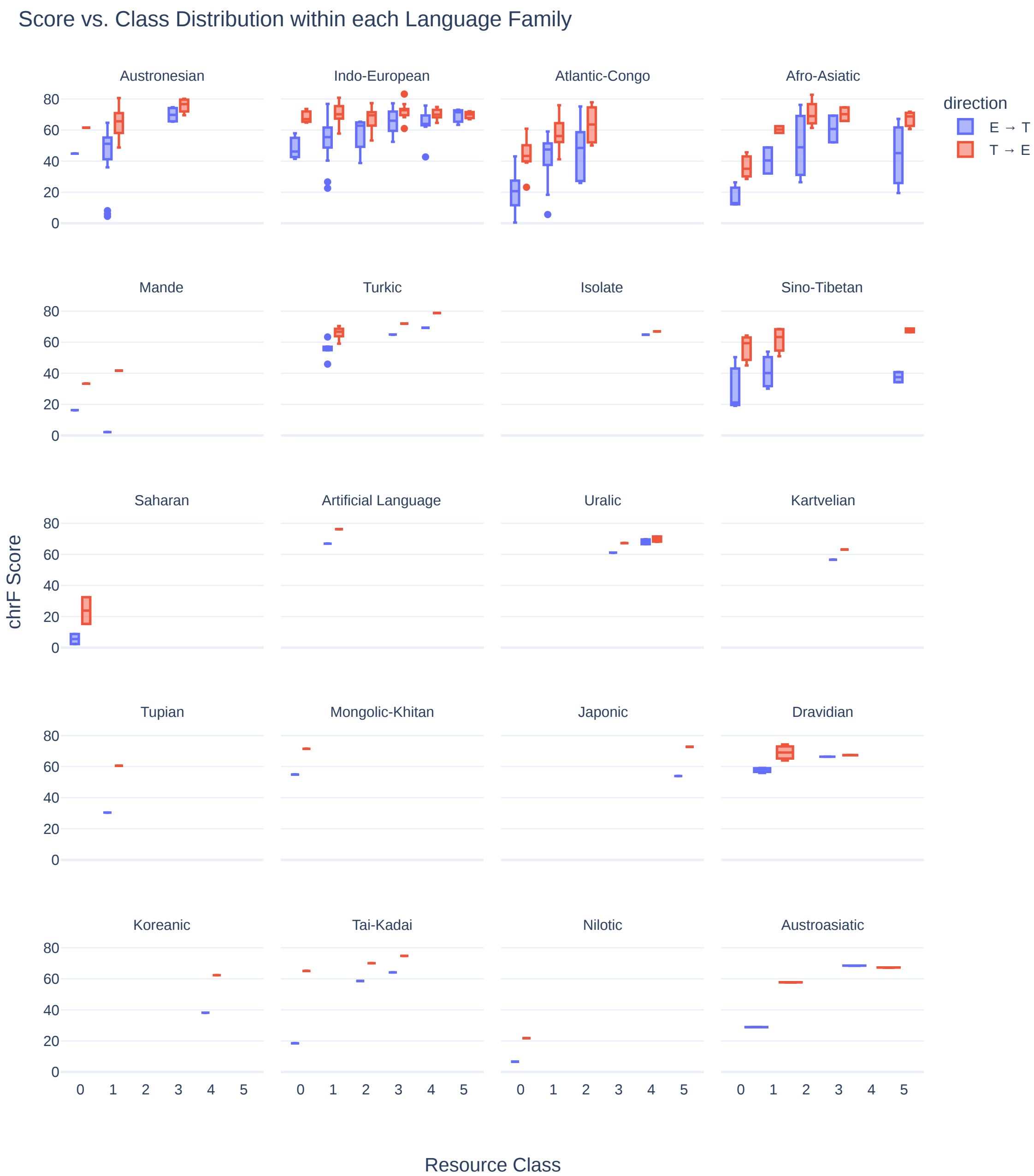


Figure 15: **Score vs. Class Distribution within each Language Family.** This faceted plot details the relationship between resource class and chrF score for each language family individually. A positive trend, where higher scores are associated with higher resource classes, is visible within several major families like Indo-European and Afro-Asiatic. The plot also highlights data sparsity, as many families (e.g., Mande, Saharan, Nilotic) contain languages in only one or two resource classes. The performance gap between the two translation directions persists even when controlling for class within a family.

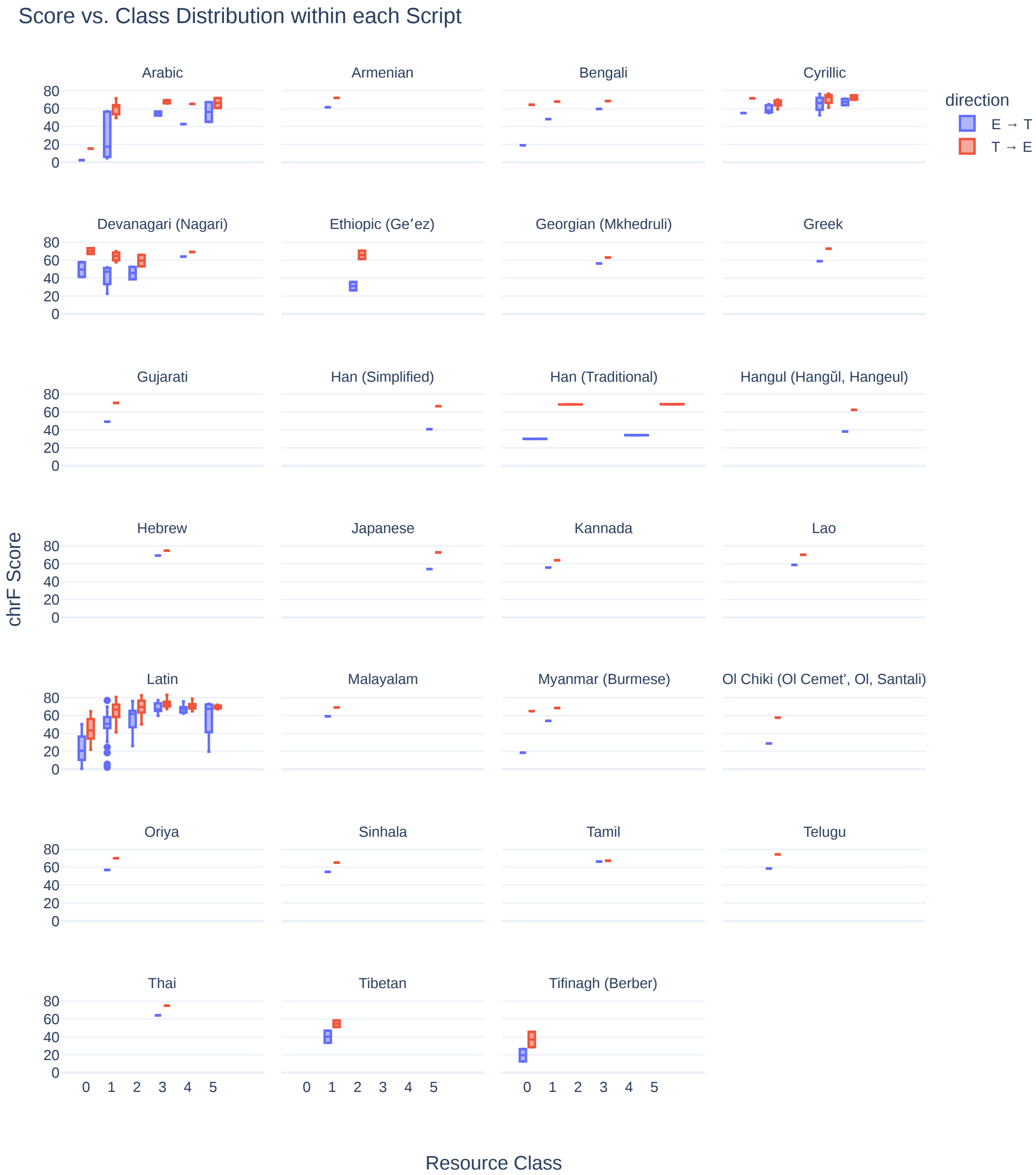


Figure 16: **Score vs. Class Distribution within each Script.** This faceted plot shows the relationship between resource class and chrF score for each writing system. The Latin script subplot contains the most data across all resource classes and most clearly demonstrates the positive correlation between class and score. For many other scripts, such as Arabic and Devanagari, the data is concentrated in the lower resource classes. This visualization confirms that the relationship between script and score is highly confounded with resource availability.

M Full Table of Model Performances

Run ID	Avg Score (Answer)	Avg Score (Explanation)	Avg Score (Total)	p-value (Total)
Gemini-2.5-pro (baseline)	0.385	0.520	0.443	N/A
OpenAI-o4-mini (baseline)	0.193	0.332	0.256	N/A
GPT-5 (baseline)	0.332	0.532	0.420	6.75×10^{-19}
Gemini-2.5-pro (guided)	0.392	0.537	0.454	2.11×10^{-1}
OpenAI-o4-mini (guided)	0.181	0.339	0.250	4.04×10^{-1}
Gemini-2.5-pro (w/ grammar agent)	0.383	0.533	0.448	5.50×10^{-1}
Gemini-2.5-pro (Single agent, 1st round)†	0.383	0.522	0.444	N/A
Gemini-2.5-pro (Single agent, 2 rounds)	0.392	0.554	0.463	1.31×10^{-2}
Gemini-2.5-pro (Single agent, 3 rounds)	0.397	0.553	0.465	7.37×10^{-3}
Gemini-2.5-pro (Single agent, 4 rounds)	0.404	0.563	0.473	4.48×10^{-4}
Gemini-2.5-pro (Single agent, 5 rounds)	0.407	0.569	0.478	7.08×10^{-5}
Gemini-2.5-pro (Single agent, 6 rounds)	0.409	0.567	0.478	1.02×10^{-4}
OpenAI-o4-mini (Single agent, 1st round)†	0.180	0.344	0.253	N/A
OpenAI-o4-mini (Single agent, 2 rounds)	0.191	0.357	0.264	2.40×10^{-1}
OpenAI-o4-mini (Single agent, 3 rounds)	0.192	0.367	0.269	6.69×10^{-2}
OpenAI-o4-mini (Single agent, 4 rounds)	0.197	0.357	0.267	1.30×10^{-1}
OpenAI-o4-mini (Single agent, 5 rounds)	0.199	0.371	0.274	1.20×10^{-2}
OpenAI-o4-mini (Single agent, 6 rounds)	0.198	0.378	0.276	4.29×10^{-3}
Gemini-2.5-pro (MoA, 1st round)†	0.389	0.540	0.453	N/A
Gemini-2.5-pro (MoA, R=0, (2 rounds))	0.398	0.556	0.466	1.49×10^{-2}
Gemini-2.5-pro (MoA, R=1, (3 rounds))	0.410	0.573	0.480	7.74×10^{-5}
Gemini-2.5-pro (MoA, R=2, (4 rounds))	0.417	0.569	0.481	1.08×10^{-4}
Gemini-2.5-pro (MoA, R=3, (5 rounds))	0.418	0.581	0.488	1.06×10^{-5}
Gemini-2.5-pro (MoA, R=4, (6 rounds))	0.421	0.579	0.489	1.50×10^{-5}
OpenAI-o4-mini (MoA, first round)†	0.187	0.344	0.257	N/A
OpenAI-o4-mini (MoA, R=0 (2 rounds))	0.325	0.491	0.397	2.70×10^{-16}
OpenAI-o4-mini (MoA, R=1 (3 rounds))	0.359	0.513	0.427	2.83×10^{-18}
OpenAI-o4-mini (MoA, R=2 (4 rounds))	0.366	0.531	0.438	2.12×10^{-20}
OpenAI-o4-mini (MoA, R=3 (5 rounds))	0.384	0.537	0.451	1.83×10^{-20}
OpenAI-o4-mini (MoA, R=4 (6 rounds))	0.392	0.543	0.457	1.07×10^{-20}

Table 8: Summary of agent performance, showing average scores of "answer", "explanation" and the combined total score. Each row represents a unique experimental setting. For the results with multiple rounds, the name denotes the model used in the final layer (i.e, the final solution is generated by it). The p-value is calculated with paired Student's t-test, comparing the model with the baseline model of the same family. The rows marked with a dagger (†) means that its setting is equivalent to the baseline, and therefore the score differences demonstrate model stochasticity.

N Scores Categorized by Language Family and Problem Type

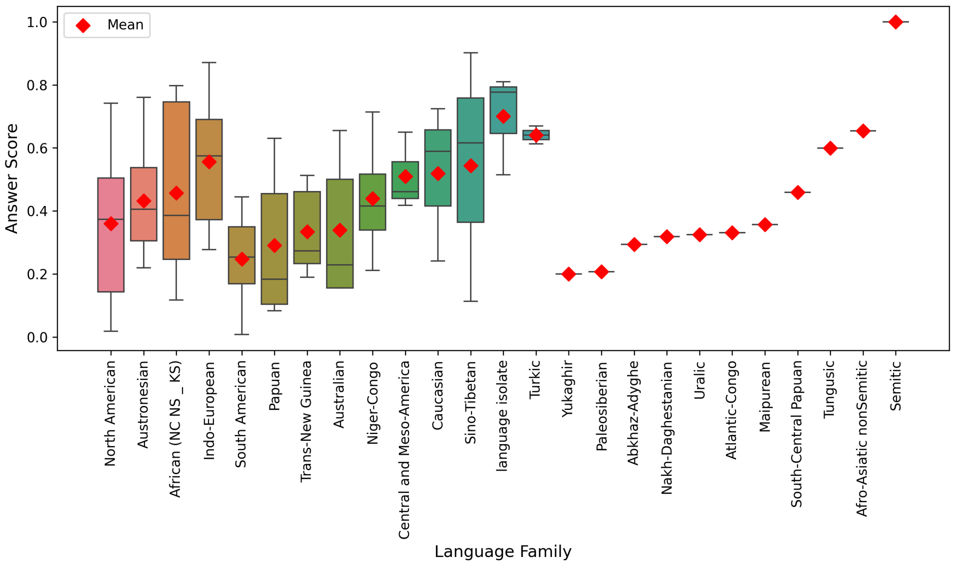


Figure 17: Distribution of Scores by Language Family.

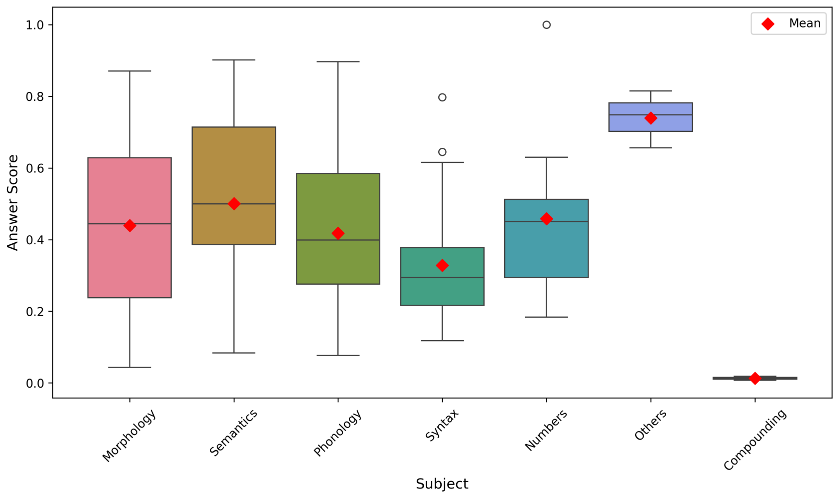


Figure 18: Distribution of Scores by Subject.

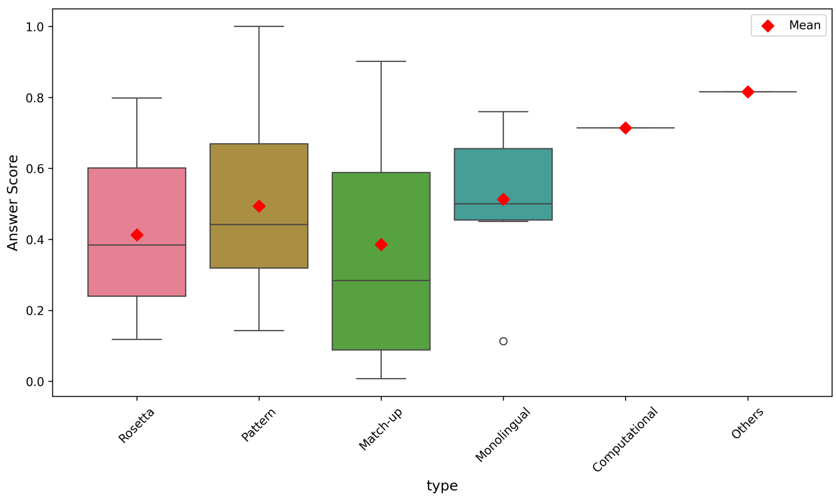


Figure 19: Distribution of Scores by Problem Type.

O Correlation between Answer Scores and Explanation Scores

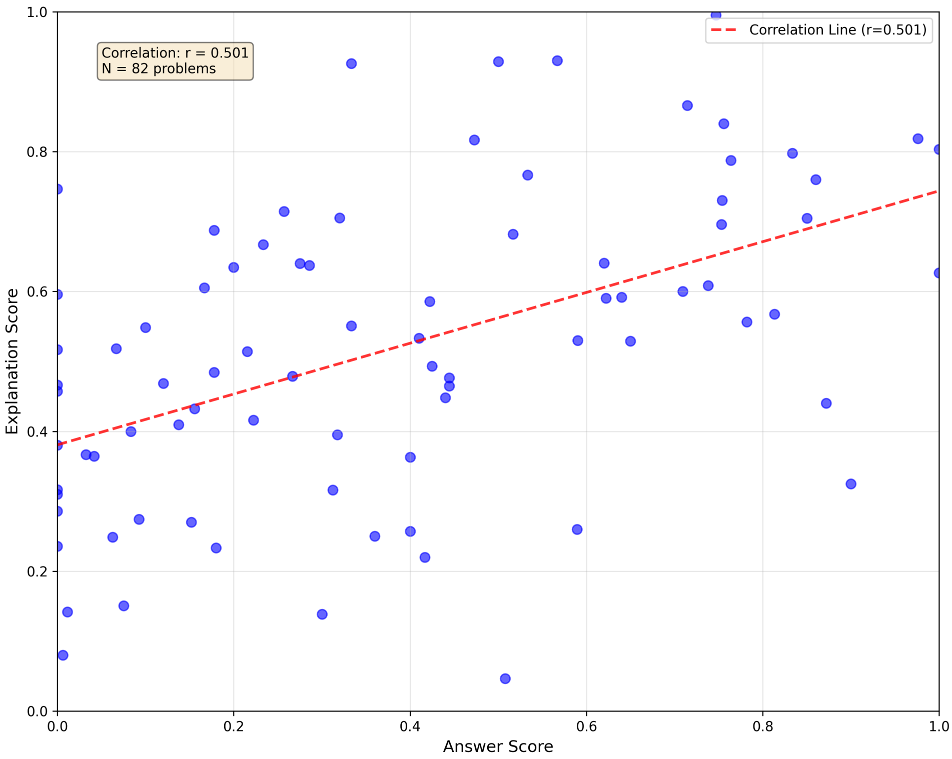


Figure 20: Correlation between Answer Scores and Explanation Scores.

跨使用者協同與序列建模的推薦系統
Cross-user Collaborative and Sequential Modeling for Recommendation

Qiao-Ying He, Yi-En Chen, Kuan-Yu Chen
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
qiaoyinghe.cs@gmail.com, yechen.research@gmail.com, kychen@mail.ntust.edu.tw

摘要

多行爲推薦系統透過引入輔助行爲，有效緩解了目標行爲的稀疏性問題。現有方法大致可分爲兩類：序列模型能夠捕捉個體的時序動態，但往往忽略跨用户的協同資訊；圖模型則能挖掘用户間的協同模式，卻缺乏對時序依賴的建模。爲此，本文提出整合序列模型與圖模型：前者專注於建模用户行爲序列的時序依賴，後者則統計並挖掘跨用户的行爲路徑。最終，通過融合兩者的預測結果，實現更爲精確的推薦效能。在兩個電商資料集 Taobao 和 RetailRocket 上，整合方法相比最強基準 MB-STR，HR@10 和 NDCG@10 均提升約 1%。實驗結果表明，即使在強大的序列模型基礎上，引入跨用户協同訊息仍能帶來穩定的性能提升。

Abstract

Multi-behavior recommendation leverages auxiliary behaviors to effectively alleviate the sparsity of target behaviors. Existing approaches can be broadly categorized into two paradigms: sequential models that capture individual temporal dynamics but often omit cross-user information, and graph-based models that mine collaborative patterns yet lack temporal dependency modeling. To address these limitations, this paper proposes an integrated approach that combines sequential and graph modeling: the former focuses on learning temporal dependencies within user behavior sequences, while the latter captures cross-user behavior paths. By fusing the predictions from both components, the method achieves more accurate recommendations. Experiments on two e-commerce datasets, Taobao and RetailRocket, show that the integrated model outperforms the strong baseline MB-STR by about 1% in both HR@10 and NDCG@10. These results indicate that incorporating cross-user collaborative information consistently improves performance, even on top of strong sequential models.

關鍵字：多行爲推薦、序列模型、圖模型、稀疏性

Keywords: Multi-Behavior Recommendation, Sequential Models, Graph-based Models, Sparsity

1 引言

推薦系統在現代電子商務平台中扮演著關鍵角色，透過預測用户偏好來緩解資訊過載問題。在實際應用中，用户與項目的互動呈現多行爲特性，包括點擊、收藏、加入購物車和購買等多種行爲類型。這些不同類型的行爲反映了用户不同層次的興趣和意圖：點擊代表初步興趣，加入購物車顯示購買意願，而購買則是最終的目標行爲。多行爲推薦系統旨在利用這些豐富的輔助行爲來改善對稀疏目標行爲的預測。

近年來，多行爲推薦研究主要沿著兩個方向發展。第一個方向是序列建模方法，如 MB-STR (Yuan et al., 2022)、DMT (Gu et al., 2020) 等，這些方法將用户的多行爲交互視爲時序序列，利用 RNN 或 Transformer 架構捕捉行爲之間的時序依賴關係。這類方法能夠精確地爲個人行爲的演化過程建模，理解用户興趣的動態變化。然而，它們僅關注單一用户的歷史序列，忽略了其他用户的行爲模式可能提供的有價值資訊。

第二個方向是基於圖結構的方法，包括利用圖神經網路的 MBGCN (Jin et al., 2020)、MB-GMN (Xia et al., 2021)，以及基於模式挖掘的 BPMR (Li et al., 2024)。這些方法將用户-項目交互建模爲圖結構，通過分析圖中的連接模式來進行推薦。特別是 BPMR，它通過統計用户-項目二部圖中的多條路徑（如「用户 A 瀏覽項目 X → 用户 B 也瀏覽項目 X → 用户 B 購買項目 Y」）來發現跨用户的行爲模式。這類方法有效利用了群體智慧，但在處理

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 230–236
November 20-22, 2025 ©2025 Association for Computational Linguistics

時序資訊和個人化建模方面存在不足。

這兩種範式各有優勢但也存在明顯的互補性。序列方法擅長建模個人化行爲模式，通過捕捉用户歷史交互的時序演化來理解個體偏好的動態變化；而跨用户方法擅長挖掘協同訊息，通過分析群體行爲模式來發現集體智慧和用户間的相似性。一個理想的推薦系統應該能夠同時考慮這兩個維度：既要精確建模個人行爲的時序動態，又能有效利用跨用户的協同模式。

有鑑於此，本文提出一個整合框架，將個人序列建模與跨用户模式挖掘結合。我們的主要貢獻包括：

- 整合框架：設計了一個簡潔的融合機制，通過 MLP 處理 MB-STR 序列表徵與 BPMR 跨用户統計的串接結果，實現兩種訊息源的整合。
- 實驗驗證：在兩個眞實資料集上進行實驗，驗證了即使採用簡單的融合策略，整合方法仍能帶來穩定的性能提升，爲多行爲推薦中不同訊息源的結合提供實證支持。

2 相關研究

2.1 序列推薦

序列推薦系統從用户的歷史交互序列中學習時序模式以預測下一個項目。早期研究使用馬可夫鏈方法建模項目間的轉換關係 (Rendle et al., 2010)。隨著深度學習的發展，循環神經網路首先被應用於序列建模，如 GRU4Rec (Hidasi et al., 2015) 利用 GRU 網路處理點擊序列，展現了深度學習在捕捉複雜序列依賴上的潛力。卷積神經網路也被用於序列推薦，Caser (Tang and Wang, 2018) 將序列嵌入視爲圖像，通過卷積操作提取局部序列模式。Transformer 架構的引入爲序列推薦帶來突破性進展。SASRec (Kang and McAuley, 2018) 採用單向自注意力機制，有效捕捉項目間的長程依賴關係。BERT4Rec (Sun et al., 2019) 進一步引入雙向注意力和掩碼語言模型，通過預訓練提升序列表徵質量。TiSASRec (Li et al., 2020) 考慮時間間隔訊息，增強了時序建模能力。這些基於注意力的方法相比 RNN 具有更好的並行性和長程依賴建模能力。在多行爲場景下，序列建模面臨新的挑戰。DMT (Gu et al., 2020) 使用多任務學習框架同時建模多種行爲序列，但採用固定的行爲模式。DIPN (Guo et al., 2019) 通過層級注意力網路建模行爲間關係，但仍在行爲層級聚合。MB-STR (Yuan et al., 2022) 提出多行爲 Transformer 層，通過行爲特定的注意力機制和位置編碼捕捉異質行爲序列的細粒度依賴。然而，這些方法都侷限於單一用户的歷史序列，未能利用跨用户的協同訊息。

2.2 基於圖的多行爲推薦

基於圖的多行爲推薦將用户-項目交互建模爲異質圖結構，利用圖學習技術進行推薦。早期方法通過矩陣分解處理多行爲數據，如 CMF (Singh and Gordon, 2008) 同時分解多個行爲矩陣並共享嵌入。隨著圖神經網路的發展，研究者開始利用 GNN 的強大表徵學習能力處理多行爲圖結構。圖神經網路方法通過消息傳遞機制聚合鄰居資訊。MBGCN (Jin et al., 2020) 在統一的用户-項目圖上進行圖卷積，並通過行爲特定的圖學習行爲語義。GHCF (Chen et al., 2021) 設計層級圖卷積網路，分別建模每種行爲的貢獻度。MB-GMN (Xia et al., 2021) 採用元學習網路自適應地學習不同行爲的權重。MGNN (Zhang et al., 2020) 利用多層網路結構同時學習共享和行爲特定的嵌入。最近的 MBCGCN (Cheng et al., 2023) 通過級聯圖卷積顯式建模行爲間的依賴關係。除了神經網路方法，統計模式挖掘提供了另一種視角。BPMR (Li et al., 2024) 不同於 GNN 的隱式表徵學習，通過統計用户-項目二部圖中的多跳路徑（如 view→view→purchase）來挖掘跨用户的行爲模式。這種方法提供了可解釋的統計特徵，能夠直接量化不同行爲模式對推薦的影響。相比 GNN 需要多層傳播可能導致的過平滑問題，BPMR 通過顯式的路徑計數保留了細粒度的行爲轉換訊息。儘管圖方法有效利用了跨用户資訊，但在處理時序動態方面存在不足。現有工作要麼完全忽略時間訊息，要麼僅將其作爲額外特徵，未能充分建模用户興趣的時序演化。本文旨在結合序列建模與圖結構方法的優勢，同時捕捉個人時序動態和跨用户協同模式。

3 預備知識

3.1 問題定義

在這一節中，我們將對多行爲推薦問題進行形式化定義。

首先，我們考慮一個存在多種交互行爲的推薦場景。該場景包含一個用户集合 $\mathcal{U} = \{u_1, u_2, \ldots, u_{|\mathcal{U}|}\}$ 和一個項目集合 $\mathcal{I} = \{i_1, i_2, \ldots, i_{|\mathcal{I}|}\}$。用户與項目之間存在多種類型的交互行爲，我們將這些行爲類型定義爲一個集合 $\mathcal{B} = \{b_1, b_2, \ldots, b_{|\mathcal{B}|}\}$。其中，$b_{|\mathcal{B}|}$ 表示我們的目標行爲（例如購買），而其餘的 $b_1, \ldots, b_{|\mathcal{B}|-1}$ 則爲輔助行爲(例如瀏覽、收藏、

加入購物車等)。

基於上述設定,每個用户的歷史行爲可以被記錄爲一個時間有序的序列。

定義 3.1 (多行爲交互序列). 對於任一用户 $u \in \mathcal{U}$,其多行爲交互序列定義爲 $S_u = \left[(i_1^u, b_1^u), (i_2^u, b_2^u), \ldots, (i_{|S_u|}^u, b_{|S_u|}^u)\right]$。序列中的每一個元組 (i_j^u, b_j^u) 表示用户 u 在某個時間點對項目 $i_j^u \in \mathcal{I}$ 執行了行爲 $b_j^u \in \mathcal{B}$。整個序列 S_u 嚴格按照交互發生的時間戳升序排列。爲了在模型中進行批次處理,我們將所有用户的序列處理爲固定長度 n。若序列的原始長度超過 n,則保留最近的 n 次交互;若長度不足 n,則在序列的開頭填充特殊的 *[PAD]* 標記。

除了用户自身的行爲序列,用户與項目之間的協同關係也蘊含了豐富的訊息。我們可以將所有交互行爲建模爲一個用户-項目圖,並從中提取高階的關聯模式。

定義 3.2 (跨用户行爲模式). 給定一個由所有交互 (u, i, b) 構成的用户-項目圖 $G = (\mathcal{U} \cup \mathcal{I}, E)$,一個長度爲 l 的行爲模式路徑可以表示爲 $p = n_1 \xrightarrow{b_1} n_2 \xrightarrow{b_2} \ldots \xrightarrow{b_l} n_{l+1}$,其中節點 $n_i \in \mathcal{U} \cup \mathcal{I}$。該路徑揭示了節點之間通過一系列特定行爲 $R = b_1 \circ b_2 \circ \ldots \circ b_l$ 所形成的關聯模式。

本文的目標是進行多行爲推薦。具體而言,給定用户 u 的定長多行爲交互序列 S_u 以及一組從圖中提取的跨用户行爲模式 P,我們的任務是預測用户 u 在下一個時間步(即 $n+1$)對任意候選項目 $i \in \mathcal{I}$ 執行目標行爲 $b_{|\mathcal{B}|}$ 的概率,記爲 $P(i, b_{|\mathcal{B}|}|S_u, P)$。

4 模型架構

本節詳述我們的模型架構,該架構整合序列建模與跨用户模式挖掘兩種互補方法。整體框架包含三個主要部分:序列行爲建模模組、跨用户模式提取模組,以及一個融合預測的模組。

4.1 序列行爲建模

在本研究中,我們採用 MB-STR (Yuan et al., 2022) 模型來進行序列行爲建模。MB-STR 包含多行爲 Transformer 層(MB-Trans)、多行爲序列模式生成器(MB-SPG)和行爲感知預測模組(BA-Pred)。

給定用户 u 的多行爲序列 $S_u = [(i_1^u, b_1^u), \ldots, (i_n^u, b_n^u)]$,我們首先僅使用項目 ID 序列來構建 Transformer 的初始輸入表徵 $\mathbf{H}^{(0)}$:

$$\mathbf{H}^{(0)} = \mathbf{E}_{\text{item}}(i_1^u, \ldots, i_n^u)$$

其中 $\mathbf{E}_{\text{item}}$ 爲項目嵌入矩陣。對應的行爲序列 $\mathcal{B}_u = (b_1^u, \ldots, b_n^u)$ 將作爲獨立的輸入,傳遞給後續的 Transformer 層,用於實現行爲感知的計算。

接著,MB-STR 採用行爲感知的多頭自注意力。對於查詢位置 i 和鍵位置 j,其對應的行爲分別爲 b_i 和 b_j。注意力分數的計算結合了內容資訊和行爲模式偏置:

$$A_{i,j} = \frac{(Q_i W_{b_i,b_j} K_j^T)}{\sqrt{d}} + P_{b_i,b_j}(j-i)$$

其中 W_{b_i,b_j} 是根據行爲對 (b_i, b_j) 選擇的特定變換矩陣。$P_{b_i,b_j}(j-i)$ 是由 MB-SPG 模組根據行爲序列 $\mathcal{B}_u$ 和相對位置 $(j-i)$ 動態生成的偏置項。

每層的輸出會通過一個行爲特定的多層感知器(BS-MLP),該模組爲每種行爲類型都維護一組獨立的參數。整體採用 L 層堆疊結構,最終得到精煉後的序列表徵 $\mathbf{H}^{(L)} \in \mathbb{R}^{n \times d}$。

4.2 跨用户的協同資訊

另一方面,我們採用 BPMR (Li et al., 2024),通過統計用户-項目交互圖中的行爲路徑來捕捉跨用户的協同資訊。

首先,我們考慮長度 $l \in \{1, 3\}$ 的行爲模式路徑。長度爲 1 的路徑表示直接的用户-項目交互(如 $u \xrightarrow{\text{buy}} i$),長度爲 3 的路徑則捕捉跨用户的協同模式(如 $u \xrightarrow{\text{view}} i' \xleftarrow{\text{view}} u' \xrightarrow{\text{buy}} i$)。BPMR 同時統計這兩種長度的路徑數量,用於計算貝葉斯評分。

接著,我們對預定義的模式集合 $\mathcal{T}$ 中的每種模式 $R \in \mathcal{T}$ 進行統計,並計算其對數貝葉斯分數。我們將用户 u 與項目 i 之間的模式總分簡記爲 $s_{u,i}^{\text{patt}}$:

$$s_{u,i}^{\text{patt}} = \sum_{R \in \mathcal{T}} N_R^{u,i} \cdot \log \left[\frac{P(R|y=1)}{P(R|y=0)}\right]$$

其中 $y = 1$ 表示用户-項目對存在目標行爲(購買),$y = 0$ 表示不存在目標行爲。$P(R|y=1)$ 和 $P(R|y=0)$ 分別代表在有購買和無購買的情況下,觀察到模式 R 的條件概率。路徑計數 $N_R^{u,i}$ 表示用户 u 和項目 i 之間存在多少條類型爲 R 的路徑。對於長度爲 3 的路徑,計數通過交互矩陣乘積 $[\mathbf{E}^{b_1} \cdot (\mathbf{E}^{b_2})^T \cdot \mathbf{E}^{b_3}]_{u,i}$ 獲得,其中 $\mathbf{E}^b$ 是行爲 b 的用户-項目交互矩陣。

4.3 融合預測

本模組在預測分數(Logits)層級整合來自序列模組和模式模組的訊號。對於給定的用户 u 和候選項目 i,兩個模組分別產生預測分數:

- 序列行爲模型 $\hat{y}^{\text{seq}}_{u,i}$：由序列行爲建模的最終表徵 $\mathbf{H}^{(L)}$ 經過行爲感知預測模組 (BA-Pred) 計算得出。

- 跨用户協同資訊模型 $s^{\text{patt}}_{u,i}$：將預先計算的模式總分取對數。

我們將 $\hat{y}^{\text{seq}}_{u,i}$ 和 $s^{\text{patt}}_{u,i}$ 拼接成一個二維向量，並將其輸入到一個前饋神經網路（Feed-Forward Neural Network, FFNN）中，以非線性的方式學習兩者間的關係，並產生最終的預測分數：

$$\hat{y}^{\text{final}}_{u,i} = \text{FFNN}\left(\text{Concat}\left(\hat{y}^{\text{seq}}_{u,i}, s^{\text{patt}}_{u,i}\right)\right)$$

4.4 模型訓練

對於序列模型 MB-STR，我們使用遮蔽語言模型（Mask Language Model, MLM）爲訓練目標，也就是通過最小化負對數概似損失來優化序列模型參數 θ_{seq}：

$$\mathcal{L}(\theta_{seq}) = -\sum_{u\in\mathcal{U}}\sum_{k=1}^{n} m_{u,k}\log p_{\theta_{seq}}(i_k|S^{\text{m}}_u)$$

其中 $m_{u,k}$ 是遮罩標記，當位置 k 被遮蔽時爲 1，否則爲 0。S^{m}_u 表示遮蔽後的序列。

在 MB-STR 訓練完成後，我們再針對融合預測模組的 FFNN 參數進行訓練，此時 BPMR 分數也已預先計算完成，因此訓練目標爲最小化融合後的預測損失：

$$\mathcal{L}(\theta_{FFN}) = -\sum_{u\in\mathcal{U}}\sum_{k=1}^{n} m_{u,k}\log p_{\theta_{FFN}}(i_k|S^{\text{m}}_u, P_u)$$

其中 Pu 包含預計算的 BPMR 模式分數，θ_{FFN} 代表 FFNN 層的模型參數。

5 實驗

5.1 實驗設定

5.1.1 資料集

我們在兩個眞實世界的電子商務資料集上進行實驗：

- **Taobao** [1]：數據來自中國最大電商平台淘寶，時間跨度爲 2017 年 11 月 25 日至 12 月 3 日。包含瀏覽（view）、加入購物車（cart）和購買（buy）三種行爲類型。

- **RetailRocket** [2]：數據來自俄羅斯電商平台的公開資料集，收集於約 4 個月期間。包含瀏覽（view）、加入購物車（cart）和購買（buy）三種行爲類型。

[1] `https://tianchi.aliyun.com/dataset/649`

[2] `https://www.kaggle.com/retailrocket/ecommerce-dataset`

資料集的詳細統計訊息如表1所示。兩個資料集都以購買（buy）作爲目標行爲，其餘爲輔助行爲。

5.1.2 評估協議與參數設置

採用 leave-one-out 策略，將每個用户的最後一個目標行爲作爲測試集，倒數第二個作爲驗證集，其餘作爲訓練集。

我們使用兩個廣泛應用於推薦系統的評估指標：

- **Hit Ratio@K (HR@K)**：衡量前 K 個推薦項目中是否包含眞實項目的比例。HR@K 反映了推薦系統的召回能力，計算公式爲：

$$\text{HR@K} = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\mathbb{I}(\text{rank}_u \le K)$$

其中 $\mathbb{I}(\cdot)$ 是指示函數，rank_u 是眞實項目在推薦列表中的排名。

- **Normalized Discounted Cumulative Gain@K (NDCG@K)**：考慮排名位置的評估指標，對排名靠前的正確推薦給予更高權重：

$$\text{NDCG@K} = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{\mathbb{I}(\text{rank}_u \le K)}{\log_2(\text{rank}_u + 1)}$$

我們的模型是基於 PyTorch 進行實現, 參數設置如下：序列長度 $n = 50$，嵌入維度 $d = 16$，MB-STR 層數 $L = 2$，注意力頭數 $h = 2$。學習率設爲 0.001，批次大小爲 128，dropout 率爲 0.2。BPMR 考慮長度爲 1 和 3 的行爲路徑。使用 Adam 優化器進行最佳化。

5.1.3 Baselines

我們與以下三類基準方法進行比較：

1. 單行爲序列模型 **(Single-behavior Sequential Models)**：
   - GRU4Rec (Hidasi et al., 2015)：基於 GRU 的序列推薦。
   - SASRec (Kang and McAuley, 2018)：基於自註意力的序列推薦。
   - BERT4Rec (Sun et al., 2019)：基於雙向 Transformer 的序列推薦。

2. 基於圖的方法 **(Graph-based Methods)**：
   - LightGCN (He et al., 2020)：輕量級圖卷積網路。

Datasets	#User	#Item	#Interactions	Interaction Behavior Type
Taobao	48,749	39,493	1,952,931	{View, Cart, Buy}
RetailRocket	147,894	99,037	2,756,101	{View, Cart, Buy}

表 1: 資料集統計訊息。

Algorithm	**Taobao**		**RetailRocket**	
	HR@10	NDCG@10	HR@10	NDCG@10
單行爲序列模型 **(Single-behavior Sequential Models)**				
GRU4Rec	0.368	0.215	0.355	0.228
SASRec	0.390	0.249	0.178	0.108
BERT4Rec	0.254	0.171	0.350	0.229
基於圖的方法 **(Graph-based Methods)**				
LightGCN	0.039	0.021	0.041	0.024
MBGCN	0.309	0.143	0.369	0.222
MB-GMN	0.319	0.154	0.491	0.300
BPMR	0.403	0.223	0.363	0.201
多行爲序列模型 **(Multi-behavior Sequential Models)**				
MBHT	0.745	0.559	0.361	0.239
MB-STR	<u>0.775</u>	<u>0.635</u>	<u>0.777</u>	<u>0.653</u>
Ours	**0.790***	**0.689***	**0.792***	**0.702***
Improv.	1.02%	1.08%	1.02%	1.07%

表 2: 不同方法在兩個資料集上的性能比較

- MBGCN (Jin et al., 2020)：多行爲圖卷積網路。
- MB-GMN (Xia et al., 2021)：基於圖元網路的多行爲推薦。
- BPMR (Li et al., 2024)：基於行爲模式挖掘的推薦（我們的基礎模型之一）。

3. 多行爲序列模型 **(Multi-behavior Sequential Models)**：
   - MBHT (Yang et al., 2022)：多行爲超圖 Transformer。
   - MB-STR (Yuan et al., 2022)：多行爲序列 Transformer（我們的基礎模型之一）。

5.2 整體性能比較

表2 展示了所有方法在兩個資料集上的實驗結果。從結果可以觀察到我們提出的整合方法在兩個資料集上均取得最佳性能。在 Taobao 資料集上，HR@10 達到 0.790，NDCG@10 達到 0.689，相比最強基準 MB-STR 分別提升 1.02% 和 1.08%。在 RetailRocket 資料集上，HR@10 和 NDCG@10 分別達到 0.792 和 0.702，提升幅度爲 1.02% 和 1.07%。

單行爲序列模型（GRU4Rec、SASRec、BERT4Rec）表現中等，未能充分利用多行爲訊息。基於圖的方法中，LightGCN 因僅考慮單一行爲而表現較差，而多行爲圖方法（MBGCN、MB-GMN）雖有所改善但仍不及序列方法。多行爲序列模型展現明顯優勢，特別是 MB-STR 和 MBHT，證實了同時考慮多行爲和序列訊息的重要性。

在所有基礎模型中，MB-STR 表現最佳，在 Taobao 上 HR@10 達 0.775，在 RetailRocket 上達 0.777。這驗證了序列建模在捕捉用户動態偏好上的有效性。我們的方法通過引入 BPMR 的跨用户模式，因此相較於 MB-STR，我們可以有更進一步的提升。

6 結論

本研究探討了多行爲推薦中個人序列建模與跨用户協同模式的整合問題。我們提出了一個簡單的融合框架，結合 MB-STR 的序列建模能力與 BPMR 的跨用户模式挖掘。實驗結果顯示，在 Taobao 和 RetailRocket 兩個電商資

料集上，整合方法相比最強基準 MB-STR 在 HR@10 和 NDCG@10 上均取得約 1% 的穩定提升。證明了跨用户協同訊息對序列模型的補充價值。值得注意的是，即使 MB-STR 已經是表現優異的強基準，引入 BPMR 的統計模式仍能帶來改善，這驗證了我們的核心假設：個人時序動態與群體行爲模式是互補的訊息源。

未來，我們將在這份研究的基礎上，持續朝向結合序列資訊與跨用户協同資訊，提出更完善的模型方法，並以端到端（End-to-end）的方式進行訓練，讓模型不僅在訓練上更有效率，也能夠更進一步地提升推薦系統的任務成效。

7 Acknowledgment

This work was supported by the National Science and Technology Council (NSTC) of Taiwan under Grants NSTC 112-2636-E-011-002, NSTC 112-2628-E-011-008-MY3, and NSTC 113-2640-B-002-005. Additional support was provided by the ”Empower Vocational Education Research Center” at the National Taiwan University of Science and Technology (NTUST) through the Featured Areas Research Center Program, as part of the Higher Education Sprout Project funded by the Ministry of Education (MOE), Taiwan. The authors also thank the National Center for High-Performance Computing, National Applied Research Laboratories (NARLabs), Taiwan, for providing essential computational and storage resources.

References

Chong Chen, Weizhi Ma, Min Zhang, Zhaowei Wang, Xiuqiang He, Chenyang Wang, Yiqun Liu, and Shaoping Ma. 2021. Graph heterogeneous multi-relational recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3958–3966.

Zhiyong Cheng, Sai Han, Fan Liu, Lei Zhu, Zan Gao, and Yuxin Peng. 2023. Multi-behavior recommendation with cascading graph convolution networks. In *Proceedings of the ACM Web Conference 2023*, pages 1181–1189.

Yulong Gu, Zhuoye Ding, Shuaiqiang Wang, Lixin Zou, Yiding Liu, and Dawei Yin. 2020. Deep multifaceted transformers for multi-objective ranking in large-scale e-commerce recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2493–2500.

Long Guo, Lifeng Hua, Rongfei Jia, Binqiang Zhao, Xiaobo Wang, and Bin Cui. 2019. Buying or browsing?: Predicting real-time purchasing intent using attention-based deep network with multiple behavior. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1984–1992.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Bowen Jin, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Multi-behavior recommendation with graph convolutional networks. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 659–668.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE.

Haojie Li, Zhiyong Cheng, Xu Yu, Jinhuan Liu, Guanfeng Liu, and Junwei Du. 2024. Behavior pattern mining-based multi-behavior recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 2291–2295.

Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th international conference on web search and data mining*, pages 322–330.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820.

Ajit P Singh and Geoffrey J Gordon. 2008. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 650–658.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450.

Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 565–573.

Lianghao Xia, Yong Xu, Chao Huang, Peng Dai, and Liefeng Bo. 2021. Graph meta network for multi-behavior recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 757–766.

Yuhao Yang, Chao Huang, Lianghao Xia, Yuxuan Liang, Yanwei Yu, and Chenliang Li. 2022. Multi-behavior hypergraph-enhanced transformer for sequential recommendation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2263–2274.

Enming Yuan, Wei Guo, Zhicheng He, Huifeng Guo, Chengkai Liu, and Ruiming Tang. 2022. Multi-behavior sequential transformer recommender. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1642–1652.

Weifeng Zhang, Jingwen Mao, Yi Cao, and Congfu Xu. 2020. Multiplex graph neural networks for multi-behavior recommendation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2313–2316.

Structured vs. Unstructured Inputs in LLMs: Evaluating the Semantic and Pragmatic Predictive Power in Abnormal Event Forecasting

Jou-An Chi
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan
r11142005@ntu.edu.tw

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan
shukaihsieh@ntu.edu.tw

Abstract

Large Language Models (LLMs) are increasingly applied to temporally grounded reasoning tasks, yet the role of input representation remains unclear. This paper compares structured temporal inputs, represented as Temporal Knowledge Graphs (TKGs), with unstructured captions in two settings: forecasting future events and detecting anomalies in surveillance video descriptions. To enable direct comparison, we build a unified dataset by aligning anomaly labels from UCF-Crime with caption annotations from UCA. Experiments show that unstructured captions consistently yield slightly higher scores across both tasks, but the differences do not reach statistical significance. Their trade-offs, however, differ: captions provide richer semantic cues for generation, while TKGs reduce input length, suppress noise, and enhance interpretability. These findings suggest that action-centric corpora, such as surveillance or forensic narratives, naturally lend themselves to structured representations, which can provide temporal scaffolds for timeline reconstruction and more traceable reasoning. All code, data processing scripts, and experimental results are available at our GitHub repository.[1]

Keywords: Large Language Models (LLMs), Temporal Knowledge Graphs (TKGs), Forecasting, Anomaly Detection, Structured vs. Unstructured Input, Surveillance Video Understanding

1 Introduction

Large Language Models (LLMs) have demonstrated impressive performance across a wide spectrum of natural language processing tasks, ranging from open-domain question answering to temporal reasoning (Gruver et al., 2023; Jin et al., 2023a). Yet, when these models are applied to real-world scenarios where events unfold over time—such as surveillance video understanding, event forecasting, or anomaly detection—the choice of input representation becomes crucial. The way temporal context is presented to an LLM can significantly affect its ability to generate accurate predictions or make reliable judgments (Su et al., 2024a; Zhou and Yu, 2024).

Two common approaches to representing temporal context are unstructured text and structured knowledge representations. Raw textual descriptions, such as captions or transcripts, preserve rich semantic details and contextual cues, which may benefit generative tasks. However, they are also noisy and can introduce irrelevant information that distracts the model. In contrast, Temporal Knowledge Graphs (TKGs) encode events as structured quadruples (head entity, relation, tail entity, timestamp) (Gastinger et al., 2022; Trivedi et al., 2017b), thereby distilling interactions into a more compact and less noisy form. TKGs have been widely applied in temporal reasoning tasks such as forecasting and anomaly detection (Goel et al., 2020; Lee et al., 2023a; Jin et al., 2020). They facilitate knowledge management and temporal reasoning (Ji et al., 2021; Kejriwal, 2019), but may omit subtle semantic cues available in natural language. Despite the growing interest in both representations, there remains little systematic comparison of how structured and unstructured inputs affect LLM performance across different temporal tasks.

In this work, we investigate this gap by

[1] https://github.com/lowannann/StructVsUnstruct-LLM

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 237–248
November 20-22, 2025 ©2025 Association for Computational Linguistics

asking: (1) Does structured temporal input provide advantages over unstructured input for forecasting tasks? (2) How does temporal context—whether structured or unstructured—impact anomaly detection tasks? We evaluate LLMs on two settings using a fine-grained surveillance video dataset that combines anomaly labels from UCF-Crime (Sultani et al., 2018) with caption annotations from UCA (Yuan et al., 2023).

Our contributions are threefold. First, we propose a comparative framework that evaluates structured and unstructured inputs under two complementary temporal reasoning tasks: forecasting and anomaly detection. Second, we provide empirical evidence that unstructured captions consistently perform slightly better across both tasks, though the differences are not statistically significant. This finding suggests that LLMs may not inherently favor one representation, but that the choice between structured and unstructured inputs should depend on task demands. Finally, our results carry practical implications for applying LLMs in temporally dynamic domains, highlighting how structured formats like TKGs can support contexts where reduced input cost, transparency, or traceability are essential.

2 Related Work

KGs, TKGs, and TKG Forecasting. Knowledge Graphs (KGs) organize entities and their relations into triples $\langle h, r, t \rangle$, offering a compact and interpretable representation that supports reasoning in applications such as semantic search and question answering (Kejriwal, 2019; Ji et al., 2021). However, many real-world scenarios are inherently temporal. To capture evolving dynamics, Temporal Knowledge Graphs (TKGs) extend this structure by associating each fact with a timestamp, forming quadruples $\langle h, r, t, \tau \rangle$ (Trivedi et al., 2017a; Leblay and Chekol, 2018; Goel et al., 2020; Jin et al., 2023b). This temporal extension enables modeling sequential dependencies and facilitates downstream tasks such as forecasting and anomaly detection in time-sensitive domains. By explicitly encoding temporal order, TKGs preserve event trajectories while reducing redundancy and noise compared to free-form text.

Research on TKG forecasting (TKGF) has traditionally relied on graph-based methods, which adapt knowledge graph embedding and graph neural network (GNN) architectures to temporal settings. Examples include RE-NET and recurrent RGCN variants that propagate historical states across timesteps (Jin et al., 2020; Chang et al., 2025), as well as symbolic approaches like TLogic and Temporal ILP that induce temporal rules (Liu et al., 2022; Xiong et al., 2024). While effective, these methods often require dataset-specific tuning and struggle in sparse or noisy contexts (Ma et al., 2023; Han et al., 2021). More recently, LLM-based approaches have reframed TKG forecasting as a language modeling problem, either by integrating graph embeddings into prompts (Zhang et al., 2024b; Wang et al., 2024; Zhang et al., 2024a) or by casting historical quadruples into textual sequences for in-context learning (Lee et al., 2023a; Liao et al., 2023; Luo et al., 2024). Remarkably, even general-purpose LLMs can perform competitively with specialized graph models, suggesting that LLMs capture not only semantic cues but also structural patterns in temporal data (Lee et al., 2023a).

LLMs in Forecasting and Anomaly Detection Forecasting is a fundamental temporal reasoning task that aims to predict future events or values from historical patterns. While traditionally addressed by statistical and deep learning models, recent work has demonstrated that LLMs provide strong generalization and flexible prompting mechanisms for this task (Jin et al., 2023a; Alnegheimish et al., 2024). Approaches include zero- or few-shot prompting, fine-tuning on domain-specific datasets, and direct application of foundation models. For example, Gruver et al. (2023) and Xue and Salim (2023) showed that GPT-family models and LLaMA variants can achieve competitive results on standard benchmarks in zero-shot settings, while fine-tuned BERT-based models improved regression accuracy on structured datasets (Xue et al., 2022). These studies highlight that LLMs can encode temporal dependencies through natural language interfaces, providing a flexible alternative to specialized time-series architectures.

Anomaly detection focuses on identifying deviations from expected temporal behavior and is increasingly framed as a diagnostic test of models' temporal reasoning ability (Su et al., 2024b; Zhou and Yu, 2024). LLMs have been applied here through three main strategies: using frozen encoders for log or sensor data, fine-tuning for binary anomaly classification, and prompt-based reasoning. For instance, Dang et al. (2021) fine-tuned BERT for detecting anomalies in KPI and Yahoo datasets, while Lee et al. (2023b) evaluated few-shot and zero-shot anomaly detection on system logs. Other prompt-based methods (Zhang et al., 2023; Huang et al., 2023) demonstrated that LLMs can capture subtle irregularities in noisy or weakly labeled data. Collectively, these findings suggest that LLMs not only generalize well across forecasting and anomaly detection but also provide a unified framework for handling diverse temporal reasoning tasks.

Input Representations and Prompting Strategies for LLM The representation of temporal information critically shapes how LLMs perform reasoning over time. Structured inputs—such as KG triples or graph embeddings—encode relations explicitly, providing precision and reducing ambiguity. Studies have shown that even when entity names are replaced with arbitrary IDs, LLMs can still perform forecasting by exploiting the structural patterns alone (Lee et al., 2023a). Similarly, prompts that present historical events as discrete triples allow the model to better recognize temporal dependencies than long descriptive texts, since the latter introduce noise and redundancy (Chang et al., 2024, 2025). In contrast, unstructured inputs—such as captions or free-form text—carry richer semantic information and contextual cues, but are noisier and harder for models to consistently parse.

Despite their noisiness, unstructured representations can complement structured data by capturing semantic or pragmatic information that graphs often omit. For example, textual descriptions may highlight causal links or implicit attributes useful for reasoning about events. Prior work has shown that combining structured triples with summarized or retrieved text improves model performance by balancing precision with semantic nuance (Chang et al., 2024). In temporal question answering, GenTKGQA (Gao et al., 2024) and M3TQA (Zha et al., 2024) illustrate how textual context and graph structure can be fused to cover each other's blind spots. These results suggest that structured and unstructured inputs are not mutually exclusive but offer complementary strengths: graphs provide clarity and temporal grounding, while text introduces richness and flexibility.

We regard temporal forecasting and anomaly detection as complementary settings for evaluating how LLMs process temporally structured input. Forecasting captures whether a model can extrapolate from observed sequences to anticipate plausible next events, while anomaly detection emphasizes the ability to recognize deviations that require attention to semantic coherence, pragmatic norms, and contextual irregularities. As Zhou and Yu (2024) notes, anomaly detection serves as a particularly diagnostic probe, since it goes beyond numerical accuracy and requires models to identify exceptions and contextual shifts rather than relying on surface-level continuation. Together, these two tasks provide complementary perspectives on temporal reasoning: one oriented toward projection, the other toward sensitivity to irregularities.

In this work, we leverage the UCF-Crime Annotation (UCA) dataset, whose human-written captions offer semantically and pragmatically grounded temporal descriptions of surveillance footage. By formulating both forecasting and anomaly detection on this data, we create a unified evaluation setting that allows us to examine how LLMs interpret structured inputs (TKGs) versus unstructured inputs (captions). This dual-task design is not aimed at comparing the tasks themselves, but at using them jointly to assess how input modality shapes models' ability to internalize temporal structures and reason about events.

3 Methods

3.1 Dataset

We employs the UCF-Crime dataset (Sultani et al., 2018) and its multimodal extension, the UCF-Crime Annotation (UCA) dataset (Yuan et al., 2023). UCF-Crime contains 1,900 long surveillance videos (over 128 hours) with either normal activities or one of 13 predefined anomalous event types, such as Fighting, Robbery, Arson, Assault, and Burglary. In our setting, we define an anomaly as an event or activity within a video sequence that deviates significantly from expected normal patterns of behavior. Anomalies are inherently context-dependent, rare in occurrence, and in surveillance scenarios typically correspond to suspicious or potentially criminal actions (e.g., fighting, robbery, or arson). Following prior work on video anomaly detection, anomaly labels in our experiments are derived from benchmark annotations, where each anomalous frame is marked according to the presence of such irregular or threatening activities.

While UCF-Crime provides video-level binary anomaly labels and segment-level annotations for evaluation, it lacks natural language descriptions of visual content. To address this, the UCA dataset augments UCF-Crime with over 23,000 sentence-level captions (110 hours), each temporally aligned at 0.1-second resolution. These captions describe both normal and anomalous events in detail, offering semantically and pragmatically rich accounts of evolving scenes. The integration of UCF-Crime and UCA yields a unified data with anomaly labels, temporal spans, and human-written descriptions, enabling us to compare structured inputs (e.g., TKG quadruples) and unstructured inputs (caption sequences) for LLM-based forecasting and anomaly detection.

Table 1 provides illustrative examples from this unified dataset, showing how video segments are paired with human-written captions, their corresponding TKG representations, and anomaly labels. This format highlights the dual structured—unstructured nature of the data, which supports systematic evaluation of LLMs across different input modalities.

3.2 Models Used

We employed two LLMs, each serving a distinct role in the experimental pipeline for forecasting and anomaly detection tasks.

GPT-4o-Mini (via OpenAI API). GPT-4o-Mini was used exclusively for extracting TKG representations from natural language captions. The model was accessed through the OpenAI API[2] with LangChain[3]'s `LLMGraphTransformer()` module, using a temperature of 0.1 to ensure deterministic triple extraction. No fine-tuning or post-processing was applied beyond temporal alignment. A closed-source model was selected for this step due to its superior performance in zero-shot structural parsing and KG extraction (Huang et al., 2024; Carta et al., 2023), thereby ensuring high-quality and reliable TKG representations that minimize confounding errors in downstream evaluations.

Mistral-large-latest (via Open Source API). All downstream inference—forecasting and anomaly detection—was conducted with the open-source `mistral-large-latest`[4] . This model was chosen for two main reasons: (1) its open-source nature ensures reproducibility and transparency, which are essential for academic research; and (2) as an instruction-tuned model, it demonstrates strong reasoning and generation capabilities across diverse NLP tasks. To maintain consistency, all runs used identical inference parameters: temperature = 0.1, top-p = 1.0, and maximum input length = 128. This setup guarantees a controlled comparison between structured (TKG-based) and unstructured (caption-based) inputs.

By separating the TKG extraction phase from the main evaluation model, we ensure that observed differences between input modalities stem from the LLM's reasoning capacity rather than inconsistencies in structural encoding quality.

3.3 Experiment 1: Forecasting

Objective. The forecasting experiment evaluates whether LLMs can generate semantically plausible next-event descriptions based on prior temporal context. Instead of predicting new triples, the task is framed as forecasting the natural language caption of a future video frame given preceding input in two forms: (1) structured TKG quadruples and (2) unstructured captions. The key goal is

[2] OpenAI API: `https://openai.com/index/openai-api/`

[3] LangChain: `https://python.langchain.com/docs/introduction/`

[4] Mistral AI: `https://docs.mistral.ai/getting-started/models/models_overview/`

Video Type	Timestamp	Caption (Text and TKG Format)	Anomalous
Arson	81.3–106	**Text:** The man walked down and tried to light a piece of paper but failed to light it. **TKG:** {[Man, WALKED_DOWN, Paper], [Man, TRIED_TO_LIGHT, Paper], [Man, FAILED_TO_LIGHT, Paper]}	False
	115.8–121.2	**Text:** The man returned to the Christmas tree and continued to light it and successfully lit it. **TKG:** {[Man, RETURNED_TO, Christmas Tree], [Man, CONTINUED_TO_LIGHT, Christmas Tree], [Man, SUCCESSFULLY_LIT, Christmas Tree]}	True
Burglary	254.4–255.8	**Text:** Another person opened the trunk, and there were several men in white hiding in the trunk. **TKG:** {[Another Person, HIDING_IN, Men In White]}	False
	256.1–350.4	**Text:** A total of five people gathered around the door and cooperated to pry it open. **TKG:** {[People, GATHERED_AROUND, Door], [People, COOPERATED_TO_PRY_OPEN, Door]}	True
Explosion	0.0–9.0	**Text:** Many cars were parked on the roadside and many people walking on the roadside. **TKG:** {[Cars, PARKED_ON, Roadside], [People, WALKING_ON, Roadside]}	False
	9.0–21.3	**Text:** An explosion occurred in a building and produced smoke, and the glass of the nearby building was shaken. **TKG:** {[Explosion, OCCURRED_IN, Building], [Explosion, PRODUCED, Smoke], [Building, SHAKEN, Glass]}	True

Table 1: Examples of aligned captions, their corresponding TKG quadruples, and anomaly labels across video types.

to assess semantic coherence and contextual appropriateness of the generated output. An overview of the pipeline is shown in Figure 1.

Input Settings. Two input conditions were tested:

- Structured (TKG $\rightarrow$ Text): Captions were converted into subject—relation—object triples with aligned timestamps. These quadruples were verbalized into structured prompt templates.
- Unstructured (Text $\rightarrow$ Text): Raw or lightly summarized captions were concatenated to form free-text temporal context, which was directly inserted into the prompt.

Prompt Design. Prompts were designed to ensure parity across conditions, differing only in input format. In both cases, the LLM was instructed to predict the most likely action immediately preceding an anomaly and to output exactly one complete sentence. Example prompt templates are shown in Figure 2 and Figure 3.

Prompted Generation. Formally, the prediction is modeled as:

$$\hat{y}_{text} = \Phi_{LLM}(P_{\mathcal{I}}), \quad \mathcal{I} \in \{TKG \rightarrow Text,\ Text \rightarrow Text\} \tag{1}$$

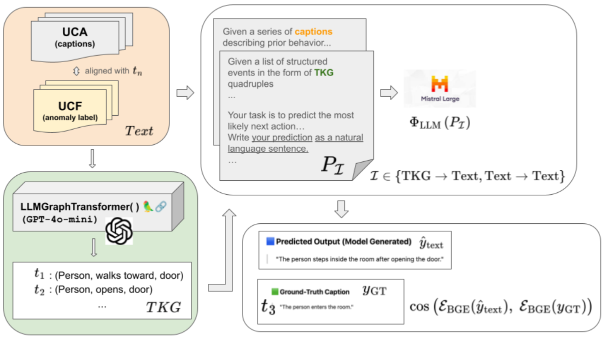


Figure 1: Pipeline of the Experiment 1: forecasting task. The model receives temporally ordered input (either structured TKGs or unstructured captions) and generates a next-frame description. The generated output is then compared against ground-truth captions to evaluate semantic alignment.

where Φ_{LLM} denotes the inference model and $P_{\mathcal{I}}$ the prompt constructed from temporal context.

Metrics Predicted sentences $\hat{y}_{text}$ were compared against human-annotated ground-truth captions y_{GT} using semantic similarity. Both sentences were encoded with the BAAI General Embedding (BGE) model, and cosine similarity was computed:

$$Similarity = \cos(\mathcal{E}_{BGE}(\hat{y}_{text}),\ \mathcal{E}_{BGE}(y_{GT})) \tag{2}$$

Cosine similarity captures paraphrastic overlap without requiring exact lexical matches, making it well-suited for evaluating free-text generation. Segment-level scores were averaged across the evaluation set to yield the final similarity metric.

3.4 Experiment 2: Anomaly Detection

Objective. The anomaly detection experiment evaluates how well LLMs identify abnormal events in surveillance video descriptions under different temporal input conditions. Given a sequence of frame-level captions, the model must judge whether the current frame is anomalous. Anomalies are defined as events that deviate significantly from expected behavioral patterns and typically correspond to suspicious or criminal actions (e.g., fighting, robbery, arson). This task probes the model's ability to reason over event coherence and detect pragmatic inconsistencies. An overview of the pipeline is shown in Figure 4.

```
[Goal]: You are given a list of structured events in the form of temporal knowledge graph (TKG) quadruples: (subject, relation, object, timestamp). These represent a subject's past actions over time.

Your task is to predict the most likely next action that the subject will perform **immediately before an abnormal event occurs**. Write your prediction as a natural language sentence.

[Input - TKG History Before Anomaly]:

T1: {[Man, WALKED_DOWN, Paper], [Man, TRIED_TO_LIGHT, Paper], [Man, FAILED_TO_LIGHT, Paper]}

T2: {[Man, RETURNED_TO, Christmas Tree], [Man, CONTINUED_TO_LIGHT, Christmas Tree], [Man, SUCCESSFULLY_LIT, Christmas Tree]}

[Constraint]:
- Predict exactly **one sentence** that describes the next likely action.
- Your output should be **one complete sentence**.

[Output - Predicted Sentence]:
```

Figure 2: Structured Input Prompt (TKG → Text) used in the forecasting task. The model is provided with a sequence of TKG quadruples representing past events and is asked to predict, in one complete sentence, the most likely next action before an anomalous event.

Prompt Design. Following the training-free strategy of Zanella et al. (2024), we prompt the LLM to assign a scalar anomaly score $a \in [0, 1]$ for each frame. Examples of each prompt are provided in Figures 5—6. The prompt is composed of three parts:

- $\mathcal{P}_S$: a system instruction framing the task as risk assessment on a 0—1 scale;
- $\mathcal{P}_F$: an output-format instruction requiring one number from a discrete set of 11 values (0.0—1.0 in steps of 0.1);
- $\mathcal{P}_C$: the temporal context, either unsummarized captions, LLM-summarized captions, or TKG quadruples:

```
[Goal]: The following is a series of natural language
captions describing the subject's behavior leading up to
an abnormal event.

Your task is to predict the most likely next action that
the subject will take right before the anomaly occurs.
The prediction should be in natural language.

[Input - Captions Before Anomaly]:

T1: The man walked down and tried to light a piece of
paper but failed to light it.

T2: The man returned to the Christmas tree
and continued to light it and successfully lit it.

[Constraint]:
- Predict exactly **one sentence** that describes the
subject's next likely action.
- Your output should be **one complete sentence**.

[Output - Predicted Caption]:
```

Figure 3: Unstructured Input Prompt (Text $\rightarrow$ Text) used in the forecasting task. The model is given a sequence of natural language captions describing prior events and is instructed to generate one complete sentence predicting the subject's next likely action before an anomaly.

$$\mathcal{C}_{temporal} \in \{\mathcal{C}_{unsummarized}, \mathcal{C}_{summarized}, \mathcal{C}_{TKG}\}. \tag{3}$$

The final prompt concatenates these components, and the LLM outputs a single anomaly score:

$$\hat{a} = \Phi_{LLM}(\mathcal{P}_S \circ \mathcal{P}_F \circ \mathcal{P}_C). \tag{4}$$

Metrics. We adopt AUC-ROC as the primary evaluation metric. Each prediction $\hat{a}$ is compared against the binary ground-truth label $a_{GT} \in 0, 1$ from UCF-Crime. AUC measures the model's ranking ability across all thresholds:

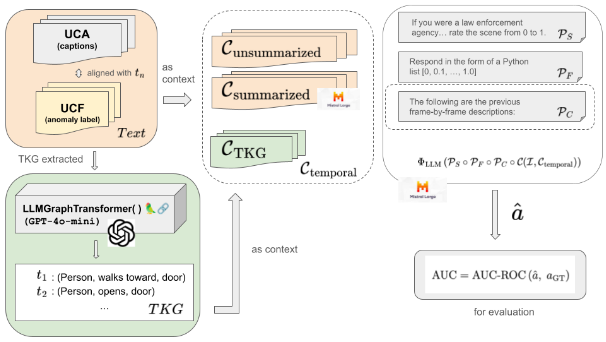


Figure 4: Pipeline of the Experiment 2: anomaly detection task. The model receives prior context in one of three forms—raw captions, summarized text, or structured TKGs—and predicts whether the current frame is anomalous. The prediction is compared against the ground-truth anomaly label for evaluation.

$$AUC = AUC - ROC(\hat{a}, a_{GT}). \tag{5}$$

Unlike accuracy, AUC is threshold-independent and reflects whether anomalies are consistently ranked above normal events. This makes it well-suited for surveillance, where operational definitions of abnormality vary. In our setting, AUC provides a robust criterion to compare how structured versus unstructured temporal inputs enable LLMs to detect deviations from normative event sequences.

4 Discussion

4.1 Research Questions

RQ1: Does structured temporal input provide advantages over unstructured input for temporal forecasting tasks? As shown in Table 2 and Table 3, the unstructured input condition yielded a slightly higher mean similarity score (0.5978) than the structured TKG input (0.5718). However, this difference was not statistically significant, as indicated by both the paired t-test ($t = 1.7259, p = 0.0978$) and the non-parametric Wilcoxon signed-rank test ($W = 90.0, p = 0.0894$). These findings suggest that unstructured input consistently produces marginally higher semantic alignment on average, though the difference does not reach statistical significance.

```
If you were a law enforcement agency, rate the scene on a scale from 0 to 1, where 0 = standard (normal) and 1 = suspicious activity (abnormal). Use the following semantic grading as guidance (examples are illustrative only; your output must still be numeric from the list below):

- 0.80–1.00 → Almost Certain / Very Likely
→ The scene is almost certainly abnormal. Example: Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism.

- 0.60–0.80 → High Suspicion
→ Clear signs of abnormality, though alternative explanations may exist. Example: pushing, running aggressively, heated argument.

- 0.40–0.60 → Uncertain / Ambiguous
→ Difficult to judge; may require human review or depend on domain context.

- 0.20–0.40 → Low Suspicion
→ Scene is mostly normal, but with minor suspicious cues. Example: unusual crowd gathering, subtle suspicious movements.

- 0.00–0.20 → Normal / Unlikely
→ Almost certainly routine daily activity.
```

Figure 5: System prompt ($\mathcal{P}_S$) used in the anomaly detection task. This instruction frames the model' s role as a law enforcement agent and asks it to assess whether the described scene is normal or suspicious on a scale from 0 (normal) to 1 (highly anomalous).

A closer inspection of forecasting outputs reveals systematic error patterns that help explain this small but non-significant gap. With TKG-based inputs, predictions often stalled at preparatory actions rather than advancing toward anomalous outcomes (e.g., anticipating ignition attempts but not the actual arson). Highly specific or low-probability events—such as an arsonist accidentally catching fire—were rarely captured, reflecting the difficulty of forecasting unexpected developments from sparse cues. The model also frequently lacked narrative progression, anchoring on earlier triplets and producing semantically plausible but stagnant outputs. Finally, forecasting performance varied by category: gradual, visually grounded events (e.g., arson, shoplifting) were more predictable than abrupt or ambiguous ones (e.g., explosions, accidents, shootings), highlighting the dependence of structured inputs on contextual richness.

```
Respond STRICTLY as a Python list containing ONE number chosen from: [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] No text, no labels, no extra spaces or characters. The response MUST begin with '[' and end with ']'.
```

```
You are given prior frame-by-frame textual descriptions as CONTEXT (no anomaly scores are provided): {context}

Use this context ONLY to understand the flow of events (e.g., who is present, what objects are involved, how actions transition over time).

⚠ IMPORTANT: The anomaly score of the CURRENT frame must be determined independently. Do NOT inherit or carry over abnormality from previous frames. Context is background information only; the final rating must reflect what is explicitly described in the CURRENT frame. Current frame description:
```

Figure 6: Prompt components for the anomaly detection task. Top: output-format prompt ($\mathcal{P}_F$), which constrains the model to return exactly one anomaly score as a Python list containing a single value between 0.0 and 1.0. Bottom: context prompt ($\mathcal{P}_C$), which provides frame-by-frame descriptions as temporal context. The context is used only to interpret event flow, while the anomaly score must be determined independently for the current frame.

These results carry important implications for the utility of structured input. TKGs offer a consistent and formal representation that abstracts away surface-level linguistic noise and encourages the model to reason based on event structure and temporal progression. This consistency may be beneficial in downstream tasks that require symbolic manipulation or multimodal alignment. By contrast, raw captions naturally carry richer lexical and syntactic cues, which directly benefit tasks emphasizing surface-level semantic similarity. While TKGs did not surpass unstructured captions in raw semantic similarity in this experiment, their representational strengths suggest potential advantages in more complex, reasoning-intensive applications.

Input Type	Cosine Similarity
Unstructured (Text → Text)	0.5978
Structured (TKG → Text)	0.5718

Table 2: Mean cosine similarity scores for structured and unstructured input conditions.

Test	Stat.	p-value
Paired t-test	$t = 1.7259$	0.0978
Wilcoxon (SR)	$W = 90.0$	0.0894

Table 3: Statistical test results comparing structured and unstructured input conditions.

RQ2: How does temporal context—whether structured or unstructured—impact LLM performance in anomaly detection tasks? We examine how temporal context—structured vs. unstructured—affects LLM anomaly detection. As shown in Fig. 7, summarized text attains AUC = 0.7817, raw text 0.7766, and TKG 0.7673. Pairwise DeLong tests (Table 4) indicate no significant differences among conditions: summarized–TKG ΔAUC = +0.014 ($p = 0.345$, 95

Qualitative error analysis reveals a few systematic behaviors. The model showed oversensitivity to ambiguous behaviors, classifying vague or cautious actions (e.g., pacing or looking around) as anomalies. Another bias appeared in action-triggered cases: attempts such as "trying to light" were flagged as anomalous even when unsuccessful.

A plausible mechanism is that TKG provides a low-noise, reference-only context. By encoding ⟨subject, relation, object, time⟩, it preserves the action backbone (who did what, when) while filtering lexical and pragmatic clutter that can nudge the model toward spurious cues. Unstructured text—especially summaries—retains fine-grained signals (e.g., negation, intensity, scene qualifiers) that occasionally help, which could explain the small numerical edge, though the average advantage remains modest. Overall, in action-centric surveillance scenes, compact structured context can achieve similar statistical performance to longer textual context while reducing token cost, offering a cost-efficient alternative when latency or context length matters.

At the same time, both tasks reveal common limitations of current LLMs for temporal reasoning: difficulty projecting narrative progression, a tendency to conflate intent with actual threat, and challenges in maintaining calibrated anomaly judgments. These findings suggest that while LLMs can leverage both structured and unstructured inputs, they still require mechanisms that better capture causal progression, distinguish ambiguous intent from concrete outcomes, and handle noisy labels.

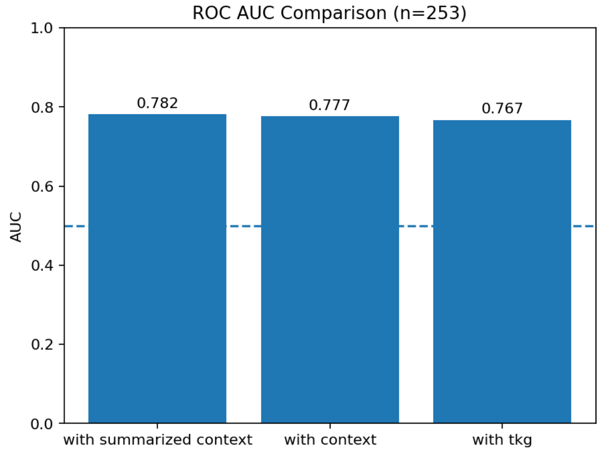


Figure 7

5 Conclusion

This study compared structured (TKG) and unstructured (caption) temporal inputs in abnormal event forecasting and anomaly detection with LLMs. Our results show that unstructured captions consistently yield slightly higher scores in both tasks, but these differences do not reach statistical significance. This finding highlights that when data are inherently action-centric—as in UCF-Crime and UCA, where human activities are described in subject—verb—object form—structured representations like TKGs provide a conceptually natural scaffold. Even when empirical gains over unstructured inputs are modest, TKGs reduce token length, enhance interpretability, and align closely with the relational structure of the data. These advantages carry practical implications for domains such as surveillance, legal, and forensic analysis, where transforming fragmented narratives into structured graphs can facilitate timeline reconstruction, highlight contradictions, and support traceable reasoning over events.

6 Limitations

Our study has several limitations. Results are based on a single open-source model (Mistral-large-latest), and may differ with other architectures or fine-tuning. Token length was not systematically explored, leaving open how

Comparison	AUC(A)	AUC(B)	ΔAUC	SE	z	p	95% CI
`raw_txt` − `tkg`	0.7673	0.7766	0.0093	0.0188	0.494	0.621	[−0.0275, 0.0461]
`sum_txt` − `tkg`	0.7673	0.7817	0.0143	0.0152	0.943	0.345	[−0.0155, 0.0441]
`sum_txt` − `raw_txt`	0.7766	0.7817	0.0051	0.0201	0.252	0.801	[−0.0344, 0.0445]

Table 4: DeLong tests for pairwise ROC AUC differences; ΔAUC = AUC(B) − AUC(A). None of the differences are statistically significant at $\alpha = 0.05$.

TKG efficiency scales under extreme long-context settings. The dataset size may also limit statistical power: unstructured inputs consistently scored slightly higher, yet differences were not significant. Finally, UCF-Crime and UCA anomaly labels may contain temporal misalignments or noise. Future work should test diverse models, larger and more varied datasets, long-context benchmarks, and improved annotations.

References

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. 2024. Large language models can be zero-shot anomaly detectors for time series? *arXiv preprint arXiv:2405.14755*.

Salvatore Carta, Alessandro Giuliani, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and Sandro Gabriele Tiddia. 2023. Iterative zero-shot llm prompting for knowledge graph construction. *arXiv preprint arXiv:2307.01128*.

He Chang, Jie Wu, Zhulin Tao, Yunshan Ma, Xianglin Huang, and Tat-Seng Chua. 2025. Integrate temporal graph learning into llm-based temporal knowledge graph model.

He Chang, Chenchen Ye, Zhulin Tao, Jie Wu, Zhengmao Yang, Yunshan Ma, Xianglin Huang, and Tat-Seng Chua. 2024. A comprehensive evaluation of large language models on temporal event forecasting. *arXiv preprint arXiv:2407.11638*.

Weixia Dang, Biyu Zhou, Lingwei Wei, Weigang Zhang, Ziang Yang, and Songlin Hu. 2021. Ts-bert: Time series anomaly detection via pre-training model bert. In *Computational Science–ICCS 2021: 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part II 21*, pages 209–223. Springer.

Yifu Gao, Linbo Qiao, Zhigang Kan, Zhihua Wen, Yongquan He, and Dongsheng Li. 2024. Two-stage generative question answering on temporal knowledge graph using large language models. *arXiv preprint arXiv:2402.16568*.

Julia Gastinger, Timo Sztyler, Lokesh Sharma, and Anett Schuelke. 2022. On the evaluation of methods for temporal knowledge graph forecasting. In *NeurIPS 2022 Temporal Graph Learning Workshop*.

Rishab Goel, Seyed Mehran Kazemi, Marcus Brubaker, and Pascal Poupart. 2020. Diachronic embedding for temporal knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3988–3995.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36:19622–19635.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2021. Robustness of temporal knowledge graph forecasting models to sparsity. In *Findings of EMNLP*.

Haoyu Huang, Chong Chen, Conghui He, Yang Li, Jiawei Jiang, and Wentao Zhang. 2024. Can llms be good graph judger for knowledge graph construction? *arXiv preprint arXiv:2411.17388*.

Shaohan Huang, Yi Liu, Carol Fung, He Wang, Hailong Yang, and Zhongzhi Luan. 2023. Improving log-based anomaly detection by pre-training hierarchical transformers. *IEEE Transactions on Computers*, 72(9):2656–2667.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. 2021. A survey on knowledge graphs: Representation, acquisition and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Di Jin, Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2020. Re-net: Reasoning over knowledge graph paths for temporal knowledge base completion. In *AAAI*.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023a. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, Shirui Pan, Vincent S. Tseng, Yu Zheng, Lei Chen, and Hui Xiong. 2023b. Large models for time series and spatio-temporal data: A survey and outlook.

Mayank Kejriwal. 2019. *Domain-specific knowledge graph construction*. Springer.

Julien Leblay and Melisachew Wudage Chekol. 2018. Deriving validity time in knowledge graph. In *Companion proceedings of the the web conference 2018*, pages 1771–1776.

Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. 2023a. Temporal knowledge graph forecasting without knowledge using in-context learning. *arXiv preprint arXiv:2305.10613*.

Yukyung Lee, Jina Kim, and Pilsung Kang. 2023b. Lanobert: System log anomaly detection based on bert masked language model. *Applied Soft Computing*, 146:110689.

Ruotong Liao, Xu Jia, Yangzhe Li, Yunpu Ma, and Volker Tresp. 2023. Gentkg: Generative forecasting on temporal knowledge graph with large language models. *arXiv preprint arXiv:2310.07793*.

Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. 2022. Tlogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4120–4127.

Ruilin Luo, Tianle Gu, Haoling Li, Junzhe Li, Zicheng Lin, Jiayi Li, and Yujiu Yang. 2024. Chain of history: Learning and forecasting with llms for temporal knowledge graph completion. *arXiv preprint arXiv:2401.06072*.

Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, and Tat-Seng Chua. 2023. Context-aware event forecasting via graph disentanglement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1643–1652.

Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. 2024a. Large language models for forecasting and anomaly detection: A systematic literature review. *arXiv preprint arXiv:2402.10350*.

Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. 2024b. Large language models for forecasting and anomaly detection: A systematic literature review.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488.

Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. 2017a. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs.

Rakshit Trivedi, Manaal Faruqui, Yann Dauphin, and Dani Yogatama. 2017b. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *ICML*.

Duo Wang, Yuan Zuo, Fengzhi Li, and Junjie Wu. 2024. Llms as zero-shot graph learners: Alignment of gnn representations with llm token embeddings. *Advances in Neural Information Processing Systems*, 37:5950–5973.

Siheng Xiong, Yuan Yang, Faramarz Fekri, and James Clayton Kerce. 2024. Tilp: Differentiable learning of temporal logical rules on knowledge graphs. *arXiv preprint arXiv:2402.12309*.

Hao Xue and Flora D Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6851–6864.

Hao Xue, Bhanu Prakash Voutharoja, and Flora D Salim. 2022. Leveraging language foundation models for human mobility forecasting. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–9.

Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. 2023. Towards surveillance video-and-language understanding: New dataset, baselines, and challenges.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. 2024. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536.

Zhiyuan Zha, Pengnian Qi, Xigang Bao, Mengyuan Tian, and Biao Qin. 2024. M 3 tqa: Multi-view, multi-hop and multi-stage reasoning for temporal question answering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10086–10090. IEEE.

Mengmei Zhang, Mingwei Sun, Peng Wang, Shen Fan, Yanhu Mo, Xiaoxiao Xu, Hong Liu, Cheng Yang, and Chuan Shi. 2024a. Graphtranslator: Aligning graph model to large language model for open-ended tasks. In *Proceedings of the ACM Web Conference 2024*, pages 1003–1014.

Ting Zhang, Xin Huang, Wen Zhao, Shaohuang Bian, and Peng Du. 2023. Logprompt: A log-based anomaly detection framework using prompts. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Wen Zhang, and Huajun Chen. 2024b. Making large language models perform better in knowledge graph completion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 233–242.

Zihao Zhou and Rose Yu. 2024. Can llms understand time series anomalies? *arXiv preprint arXiv:2410.05440*.

Embodiment in Multimodal Semantics: Comparing Sensory, Emotional, and Visual Features in Chinese Color Metaphors

Wu Yufeng
City University of Hong Kong
18, Tat Hong Avenue, Kowloon, Hong Kong
Yufenwu2-c@my.cityu.edu.hk

Liu Meichun
City University of Hong Kong
18, Tat Hong Avenue, Kowloon, Hong Kong
meichliu@cityu.edu.hk

Abstract

This study examines how sensory-motor experience, emotional valence and arousal, and visual image statistics contribute to multimodal alignment in Chinese color metaphors. Using 184 metaphorical lexemes from six basic color terms, we combined textual data from the Chinese Corpus Internet (CCI 3.0) with image sets from Baidu, embedding both with Chinese-CLIP and measuring alignment using robust pooled cosine and set-to-set Chamfer metrics. Sensory-motor ratings, especially effector exclusivity and tactile strength, correlated negatively with alignment, emotional valence showed strong positive correlations, and visual color statistics (variability, entropy) correlated positively but yielded modest generalization primarily under Chamfer. Under strict 5-fold Ridge cross-validation, emotion was the only feature group with consistently non-negative out-of-fold performance, whereas sensory ratings did not generalize. The findings indicate that affective salience and perceptual richness captured by image statistics are the principal drivers of multimodal grounding for metaphorical color words, with visual contributions emerging when alignment is evaluated many-to-many.

Keywords: Embodied cognition, Multimodal semantics, Chinese color metaphors, Text-image alignment

1 Introduction

The theory of embodied cognition proposes that word meaning is grounded in perceptual and motor experience (Barsalou, 2010; Glenberg & Kaschak, 2002; Pulvermüller, 2005). Decades of behavioral and neuroimaging research have shown that accessing word meaning can reactivate sensory-motor systems, and large-scale sensory-motor rating norms have quantified embodiment across thousands of concepts (Barsalou, 2010; Connell & Lynott, 2012; Glenberg & Kaschak, 2002; Lynott et al., 2020; Zhong et al., 2022). However, embodiment is multidimensional. Alongside sensory-motor grounding, emotion provides another pathway, with evidence that affective dimensions such as valence and arousal strongly shape lexical processing and memory (Vigliocco et al., 2009; Xu et al., 2022). Vision adds yet another layer: beyond whether a concept has visual attributes, measurable image statistics such as color entropy and variability can influence semantic representation and multimodal alignment (Jonauskaite et al., 2020; Palmer & Schloss, 2010; Radford et al., 2021; Vigliocco et al., 2009).

Despite these advances, it remains unclear how sensory, emotional, and visual factors compare in their relative contribution to cross-modal semantics, particularly in metaphorical language. Color metaphors in Chinese provide an ideal testing ground: they are rich in cultural meanings, widely represented in textual and visual data, and closely tied to both perceptual and affective associations. For instance, *lán tú* 蓝图 'blueprint'

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 249–257
November 20-22, 2025 ©2025 Association for Computational Linguistics

conveys planning and foresight, while *lǜ mào* 绿帽 'green hat' carries strong emotional connotations of betrayal.

This study therefore asks: Which embodied dimensions provide reliable signal, correlationally and out-of-fold, for alignment between linguistic and visual representations of Chinese color metaphors? To address this, we constructed a multimodal dataset of 184 color-derived lexemes, combining textual contexts from the Chinese Corpus Internet (CCI 3.0) (Liangdong Wang et al., 2024) with images retrieved from Baidu Image. Using Chinese-CLIP embeddings, we measured text–image alignment with robust pooled cosine and set-to-set Chamfer metrics, and integrated three types of features: sensory-motor ratings, emotional ratings, and image-based visual statistics. Through correlation and strict 5-fold Ridge cross-validation, we assessed both association and out-of-fold predictive power of these dimensions, including ΔR^2 contrasts to test incremental contributions, aiming to clarify the relative roles of sensory, affective, and visual factors under robust pooling and set-to-set alignment metrics, and to quantify their generalization with cross-validated models.

2 Literature review

2.1 Sensory Experience and Embodied Semantic Representations

Embodied cognition theory posits that semantic representations are partly grounded in past sensory-motor experiences, and activating word meaning will (re)engage perceptual and motor systems. A wealth of evidence supports this view:

Actions or percepts congruent with language facilitate conceptual processing. For example, compatibility between a described action and a required movement speeds comprehension. Conversely, switching between modalities incurs a processing cost in both purely perceptual tasks and conceptual tasks about perceptual properties. Such cross-modal interference suggests coupling between semantic processing and modality-specific perceptual processing. Classic studies demonstrating the action-sentence compatibility effect support this idea(Glenberg & Kaschak, 2002).

Words with action- or perception-related meanings elicit modality-specific activation in motor and sensory cortices. For instance, action words like kick or lick somatotopically activate corresponding motor regions for legs or tongue. Likewise, visual or auditory words activate occipital or temporal sensory areas (Binder & Desai, 2011; Hauk et al., 2004). Such findings reveal a systematic overlap between conceptual and perceptual brain networks, consistent with partially "embodied" semantic representations.

Large datasets have quantified the perceptual and action associations of words. In English, modality exclusivity norms rate the strength of a concept's association with five senses (and action effectors), enabling quantification of a word's "embodiment footprint". Early work by Lynott & Connell (2020) collected ratings for hundreds of concepts across modalities. Recently, the Lancaster Sensorimotor Norms provide 11-dimensional ratings (6 sensory modalities and 5 action effectors) for ~40,000 words. These norms explain differences in concreteness, category structure, and memory advantages for certain concepts. In Chinese, a systematic database of sensory-action ratings for nouns has also been developed (Zhong et al., 2022), offering modality strengths (visual, auditory, tactile, gustatory, olfactory, etc.) and bodily effectors for each word. Such resources allow researchers to characterize a word's embodied profile in multiple modalities.

Overall, the "degree of embodiment" of a concept can be operationalized via multi-modal strength, dominant modality, or modality exclusivity. These indicators correlate with concreteness and also predict imageability, memorability, and even the topology of semantic networks (Barsalou, 2010; Binder et al., 2016; Lynott et al., 2020). In multimodal tasks like image-text retrieval, incorporating sensory-motor features can complement abstract distributional vectors, improving cross-modal alignment and model interpretability. Studies have found that adding modality-specific information (e.g. visual or motor features) to word embeddings enhances performance on cross-modal matching and provides more human-interpretable alignments (Lynott et al., 2020; Shutova et al., 2016).

2.2 Emotional Dimensions and Embodied Cognition

Emotion is another key axis of embodied experience that shapes semantic representation. The classic valence-arousal model (Russell, 1980) describes emotions in a two-dimensional space (valence: positive-negative, and arousal: high-low activation). These affective dimensions are tightly coupled with attention, memory, and decision-making processes:

Psychological and neural evidence: Emotional valence and arousal modulate cognitive processing speed, memory retention, and selective attention. Positively or negatively valenced words can be processed more quickly depending on context, and high-arousal content tends to be remembered in greater detail (Kensinger, 2009). Neuroimaging and lesion studies reveal distinguishable neural signatures for valence and arousal—for example, the amygdala and ventromedial prefrontal cortex track affective intensity, while regions of prefrontal and cingulate cortex differentiate positive vs. negative valence ((Lindquist et al., 2012). Such findings suggest that emotional dimensions are instantiated in the brain's affective networks, creating an "emotional fingerprint" for semantic stimuli.

Semantic and distributional evidence: A growing body of work indicates that a word's valence and arousal ratings correlate with its position in distributional semantic space and with how it aligns to visual representations. For instance, words that are highly positive or highly negative cluster distinctly in word embedding spaces, and their emotional ratings predict human judgments and memory advantages (Hollis & Westbury, 2016; Recchia & Louwerse, 2015). High-arousal words often have more and stronger associations in semantic networks, reflecting their attention-grabbing nature.

Resources in Chinese: Recent efforts have produced emotion norms for thousands of Chinese words. Xu et al. (2022) report valence and arousal ratings for over 11,000 simplified Chinese words, with analyses of gender differences in ratings. These resources enable researchers to introduce emotion features at the word or instance level in multimodal alignment tasks. For example, one can ask whether positively valenced or high-arousal words align more easily with certain image content, or if emotional congruence between caption and image boosts alignment.

Compared to concrete sensory-motor features, emotion may be more influenced by cultural and subjective factors. However, emotion strongly influences what we attend to and remember (Kensinger, 2009). In an embodied cognition framework, emotion can be seen as an encapsulation of "embodied-social experience," complementing sensory dimensions. Together, sensory and affective features jointly determine a concept's imageability, memorability, and ease of cross-modal association. In other words, a concept rich in perceptual detail and emotional salience is more likely to be vividly visualizable and easily paired with corresponding images, providing a strong signal for image-text alignment.

2.3 Visual Information and Embodiment in Multimodal Alignment

Vision is often highlighted as a dominant embodied modality. Beyond asking whether a concept "has visual attributes," researchers are examining how measurable image statistics (brightness, color entropy, color diversity, contrast, etc.) relate to semantic representations. Several lines of inquiry demonstrate the importance of visual features:

Low-level visual statistics like color and texture carry semantic and affective connotations. Colors can imply category or function (e.g. green for plants), evoke emotions (e.g. red for anger or love), or convey symbolic meanings (Jonauskaite et al., 2020; Palmer et al., 2013). For example, warm colors (reds/yellows) are often associated with positive valence and high arousal, whereas cool colors (blue hues) tend to correlate with calmer, lower-arousal feelings. Similarly, an image's color diversity and complexity can suggest "liveliness" or conceptual richness, potentially affecting interest and memorability. Thus, concrete concepts with strong visual features might also carry consistent emotional tones (e.g., a "sunset" is visually warm and often deemed pleasant).

Modern image-text models (e.g. CLIP) implicitly capture some color and contrast information, but these can still sway alignment. Research shows that certain models behave like "bag-of-words," lacking relational understanding of image content and instead relying on object presence or overall appearance. Visual factors like brightness or saturation can sometimes confound image-text similarity if not accounted for. Using perceptually uniform color spaces (such as JzAzBz) allows more consistent quantification of image attributes like mean brightness or color entropy, which can be related to language features in an interpretable way. For instance, one might find that images with extremely high brightness are harder to align with captions due to reduced contrast, or that captions with highly concrete nouns align better with images having greater color variability (indicating more objects or details). As noted by Radford et al. (2021) and follow-up analyses, certain visual properties can either facilitate or impede cross-modal matching: a richly colored, high-contrast image may provide more "hooks" for semantic alignment, whereas an overexposed image might be less distinguishable in a joint embedding space.

Visual statistics intersect with sensory and emotional dimensions. A visually striking image (e.g., with high color variance) might align better with descriptive, concrete text, effectively leveraging embodied (visual) information to improve retrieval. At the same time, visual cues also carry emotional weight—color tone can modulate perceived valence or arousal of an image, thereby affecting alignment with text that has emotional connotations. For example, an image dominated by dark, desaturated colors might align well with a negatively valenced caption (a phenomenon related to color-emotion association). Thus, visual statistics serve as both low-level perceptual evidence and high-level semantic/emotional signals. In multimodal learning, incorporating these features can enhance alignment: one study found that adding a simple colorfulness metric improved image-caption retrieval, as it captured an aspect of "visual vividness" not present in text embeddings alone (Palmer et al., 2013; Radford et al., 2021). However, certain visual extremes (e.g., extremely bright images) can reduce alignment quality by washing out distinctive features, an observation in line with human factors in perception. The key is that visual features, in concert with sensory and emotional semantic features, contribute to a concept's overall embodied signature, which in turn influences cross-modal mapping.

3 Methodology

3.1 Lexeme Selection

Textual data were drawn from the Chinese Corpus Internet (CCI 3.0) (Liangdong Wang et al., 2024), a large-scale corpus (~1,000 GB) of digital publications from Mainland China (2001-2023). From this corpus, 184 metaphorical lexemes derived from the six basic color terms (黑 hēi 'black', 白 bái 'white', 红 hóng 'red', 黄 huáng 'yellow', 蓝 lán 'blue', and 绿 lǜ 'green') were identified using the Metaphor Identification Procedure (MIP; Pragglejaz Group, 2007).

3.2 Text and Image Data

For each lexeme, up to 100 contextual sentences were extracted from CCI 3.0 and trimmed to a ±100-character window around the target word to capture its immediate context. Parallel visual data were collected from Baidu Images, with the top 100 images per lexeme retained as representative visual exemplars after basic filtering which excluding images with resolution lower than 200x200. The text and image sets are unpaired and serve as multimodal exemplars of the same lexeme rather than item-aligned pairs.

3.3 Multimodal embeddings and alignment

Texts and images are encoded with Chinese-CLIP (ViT-L/14). To reduce sensitivity to outliers and sampling noise, we aggregate the set of text embeddings and the set of image embeddings for each lexeme using robust pooling: (i) a 10% trimmed mean and (ii) the medoid (the exemplar with minimal average cosine distance). Beyond pointwise cosine between pooled vectors, we evaluate set-to-set alignment on the full cross-modal similarity matrix $S = TI^{\top}$ after L2 normalization. We report four alignment metrics used in the Results: trimmed-cosine, spherical-cosine, agreement-weighted cosine, and bi-directional Chamfer (the average of the per-text maxima and the per-image maxima in S).

3.4 Sensory and Emotional Features

Two external rating databases were used to characterize lexemes. The Chinese Noun Sensory-Motor Norms (Zhong & Zhang, 2022), which provide ratings across six sensory modalities (visual, auditory, olfactory, gustatory, tactile, interoceptive) and associated motor effectors. The Simplified Chinese Affective Lexicon (Xu et al., 2022), which provides valence and arousal ratings on 11,310 words, with gender-specific and overall averages. Lexemes were matched to these databases to obtain multidimensional sensory and emotional ratings.

3.5 Visual Features

From the collected images, low-level visual statistics were computed in the JzAzBz perceptual color space, including average luminance, color variability, color entropy, and colorfulness. These measures captured perceptual diversity and distributional properties of the lexemes' visual exemplars.

3.6 Statistical Analyses

We assess associations and predictive power in two steps that are reported in the Results. (1) Correlation. Pearson's r between alignment scores and individual features; correlations use the trimmed-cosine alignment score. (2) Predictive modeling. Ridge regression with 5-fold cross-validation; we report CV-R^2 against two baselines: MeanBaseline (zero point) and RandomBaseline (expected ≈ -1). Models are fit for single-modality and combined feature sets, and we report ΔR^2 for nested contrasts under each alignment metric to quantify incremental value.

4 Result

4.1 Correlation Analysis

feature	r	p
sens_effector_exclusivity	-0.213	0.004
sens_tactile	-0.195	0.008
sens_max_action	-0.133	0.073
sens_auditory	0.130	0.079
sens_gustatory	-0.123	0.096
sens_concreteness	0.106	0.153
sens_head	-0.086	0.246
sens_olfactory	-0.077	0.298
sens_max_sensorimotor	-0.057	0.446
sens_perceptual_mean	-0.052	0.482
sens_max_perceptual	-0.052	0.486
sens_mouth/throat	-0.048	0.516
sens_exclusivity_sensorimotor	-0.043	0.559
sens_visual	0.042	0.568
sens_interoceptive	-0.033	0.654
sens_action_mean	-0.027	0.719
sens_torso	0.011	0.887
sens_modality_exclusivity	-0.009	0.902
sens_leg/foot	0.008	0.914
sens_hand/arm	0.002	0.976

Table 1: Pearson correlations between sensory-motor features and text-image alignment.

Table 1 reports the Pearson correlation coefficients (r) and significance levels (p) between sensory-motor dimensions and text-image alignment. Overall, most features showed weak associations with alignment, but a few dimensions yielded significant or near-significant effects.

The strongest effect was found for effector exclusivity (sens_effector_exclusivity), which correlated negatively with alignment (r = -0.213, p = .004). This suggests that lexemes characterized by greater specificity in their action effectors (e.g., strongly tied to one particular body part) tended to achieve weaker text-image alignment. In other words, highly specialized motor grounding may hinder the integration of visual and linguistic representations.

A second robust result was observed for the tactile dimension (sens_tactile), which also showed a significant negative correlation (r = -0.195, p = .008). Lexemes strongly grounded in tactile experience aligned less well across modalities, likely because tactile sensations are inherently difficult to represent visually.

Several additional features displayed marginal effects. Maximum action ratings (sens_max_action) were weakly negatively correlated with alignment (r = -0.133, p = .073), while auditory strength (sens_auditory) showed a small positive correlation (r = 0.130, p = .079), both trending toward significance. Similarly, gustatory strength (sens_gustatory) trended negatively (r = -0.123, p = .096). Although modest, these findings suggest that auditory and gustatory experiences may exert limited influence on cross-modal integration.

By contrast, most other sensory indices, such as visual (r = 0.042, p = .568), olfactory (r = -0.077, p = .298), and interoceptive (r = -0.033, p = .654), showed correlations close to zero and did not approach significance. Likewise, global embodiment indices including perceptual mean (sens_perceptual_mean) and action mean (sens_action_mean) were nonsignificant, indicating that broad averages of sensory grounding do not strongly predict alignment. These results highlight that specific embodied channels, rather than overall sensory strength, are the key drivers of cross-modal variation.

feature	r	p
emo_Women_Valence_Mean	0.448	0.000
emo_Valence_Mean	0.445	0.000
emo_Men_Valence_Mean	0.444	0.000
emo_Men_Arousal_Mean	0.283	0.000
emo_Arousal_Mean	0.230	0.002
emo_Women_Arousal_Mean	0.195	0.008

Table 2: Pearson correlations between emotional features and text–image alignment.

Table 2 presents the Pearson correlations between emotional dimensions and text-image alignment. In contrast to the sensory domain, the emotional features demonstrated consistently strong and positive relationships with alignment, particularly for valence ratings.

The valence dimension emerged as the most reliable predictor. Regardless of whether ratings were drawn from men, women, or the overall mean, the correlation coefficients were nearly identical ($r \approx 0.45$, ps < .001). This indicates that lexemes with more positive affective connotations systematically aligned better across modalities.

The arousal dimension also produced significant positive effects, though the effect sizes were smaller than those for valence. Male arousal ratings showed the strongest association (r = 0.283, p < .001), followed by the overall mean (r = 0.230, p = .002) and female arousal ratings (r = 0.195, p = .008). Collectively, these findings demonstrate that emotional positivity and activation jointly facilitate multimodal alignment, with valence providing the dominant contribution.

feature	r	p
ColorVariabilityBz	0.311	0.000
Colorfulness	0.208	0.005
ColorEntropyBz	0.205	0.005
ColorVariabilityAz	0.170	0.021
HueAngle	-0.157	0.033
ColorEntropyAz	0.148	0.045
AverageColorJz	-0.120	0.105
ColorEntropyJz	0.112	0.130
ColorContrast	0.081	0.275
AverageColorBz	-0.056	0.447
AverageColorAz	0.013	0.862
ColorVariabilityJz	-0.008	0.915

Table 3: Pearson correlations between visual color features and text–image alignment.

Table 3 lists the correlations between image-based visual features and text-image alignment. Compared to the sensory and emotional results, the visual features displayed a more mixed pattern, with some robust positive predictors alongside negative or nonsignificant effects.

The most prominent predictor was color variability along the Bz dimension (ColorVariabilityBz), which correlated moderately and positively with alignment (r = 0.311, p < .001). Two additional features—colorfulness (r = 0.208, p = .005) and color entropy in the Bz dimension (ColorEntropyBz, r = 0.205, p = .005)—also showed significant positive correlations. Together, these results suggest that lexemes whose associated images contain richer and more varied color distributions tend to achieve better cross-modal alignment.

More modest but still significant effects were found for color variability (r = 0.170, p = .021) and color entropy (r = 0.148, p = .045) in the Az dimension. These indicate that diversity in the color distribution, even along secondary axes, can enhance semantic-visual consistency.

On the other hand, some features exhibited negative or null associations. Hue angle (HueAngle) was significantly negatively correlated with alignment (r = -0.157, p = .033), implying that large deviations in hue may disrupt the semantic fit between text and images. Average lightness (AverageColorJz) showed a nonsignificant negative trend (r = -0.120, p = .105), while other mean color measures (e.g., AverageColorAz, AverageColorBz) and contrast did not contribute meaningfully (|r| < .1, ns).

4.2 Regression analysis

We assess generalization with 5-fold Ridge CV-R^2 under four alignment metrics (trimmed cosine, spherical cosine, agreement-weighted cosine, and set-to-set Chamfer). Results are summarized in Table 4. Among single-modality models, Emotion is the only group that achieves positive out-of-fold performance on the cosine family, reaching its best value with trimmed cosine (CV-R^2=0.05). Visual features alone do not surpass the mean predictor on cosine metrics but become informative when alignment is evaluated at the set level: under Chamfer the visual model attains CV-R^2=0.04. Sensory norms fail to generalize on all metrics (CV-$R^2 \leq 0$), consistent with their limited predictive value in this task.

set	cosine_trimmed	cosine_spherical	cosine_agreew	chamfer_bi
MeanBaseline	0.00	0.00	0.00	0.00
Random Baseline	-1.02	-1.10	-0.96	-0.95
sensory	-0.03	-0.05	-0.05	-0.19
emotion	0.05	0.02	-0.01	-0.07
visual	-0.05	-0.05	-0.06	0.04
sensory+emotion	0.05	0.01	-0.01	-0.18
sensory+visual	-0.03	-0.04	-0.06	-0.06
emotion+visual	0.00	-0.03	-0.03	0.01
all_three	0.02	-0.01	-0.02	-0.05

Table 4: Cross-validated R^2 values for regression models with different feature sets.

Multi-modality patterns reinforce these observations. Combining Sensory+Emotion does not improve over Emotion on trimmed cosine (0.05 vs. 0.05) and remains near zero or negative elsewhere, indicating that any apparent gains are carried by the emotional dimensions. Emotion+Visual is near zero on cosine metrics but turns positive under Chamfer (0.01). The full model does not outperform the best single or two-way combinations (e.g., −0.05 on Chamfer). Considering nested contrasts clarifies the incremental value: under Chamfer, adding Visual to Emotion yields $\Delta R^2(\mathrm{EV} - \mathrm{E}) = +0.08$, whereas the same addition is non-beneficial on cosine metrics (−0.05/−0.05/−0.02 for trimmed/spherical/agree-weighted). Thus, affective salience is the most reliable cross-modal signal, while perceptual statistics contribute a small but robust additional component specifically in the many-to-many matching regime captured by Chamfer. Negative CV-R^2 values are reported relative to the mean-predictor zero point and indicate poorer-than-mean out-of-fold prediction.

Table 5 reports nested contrasts that isolate the incremental value of each modality. Adding Emotion to Sensory consistently improves performance on all alignment metrics (SE−S: +0.07 trimmed, +0.06 spherical, +0.04 agree-weighted, +0.01 Chamfer), confirming that the gains attributed to the SE model are carried by the emotional dimensions. The reverse contrast is zero or negative (SE−E: 0.00/−0.01/0.00/−0.11), indicating that Sensory contributes no unique information beyond Emotion and can even harm generalization under set-to-set evaluation.

contrast	cosine_trimmed	cosine_spherical	cosine_agreew	chamfer_bi
SE - S	0.07	0.06	0.04	0.01
SE - E	0.00	-0.01	0.00	-0.11
SV - S	0.00	0.01	-0.01	0.13
SV - V	0.02	0.01	0.01	-0.10
EV - E	-0.05	-0.05	-0.02	0.08
EV - V	0.04	0.02	0.03	-0.03
ALL - SE	-0.03	-0.03	-0.02	0.13
ALL - SV	0.05	0.02	0.03	0.01
ALL - EV	0.02	0.01	0.01	-0.06

Table 5: Incremental ΔR^2 comparisons across feature set combinations.

The behavior of Visual depends on the alignment regime. When alignment is measured at the set level, Visual augments Sensory (SV−S: +0.13 under Chamfer), whereas adding Sensory to Visual degrades performance in the same regime (SV−V: −0.10), again pointing to the limited utility of generalized sensory norms. Critically, the Emotion+Visual contrast shows that Visual provides a small but reliable addition over Emotion only for Chamfer (EV−E: +0.08), while the cosine metrics yield non-beneficial or slightly negative increments (−0.05/−0.05/−0.02). Thus, perceptual statistics contribute additively in the many-to-many matching setting captured by Chamfer, but not in pointwise cosine alignment.

Three-way models mirror these patterns. Relative to SE, adding Visual produces a noticeable improvement under Chamfer (ALL−SE: +0.13) but not under the cosine metrics (−0.03/−0.03/−0.02). Relative to SV, adding Emotion yields small positive increments for the cosine metrics (+0.05/+0.02/+0.03) and only a negligible change for Chamfer (+0.01). In contrast, augmenting the EV model with Sensory is consistently unhelpful (ALL−EV: +0.02/+0.01/+0.01 but −0.06 under Chamfer). Taken together with Table 4, these contrasts substantiate a clear hierarchy: Emotion is the only modality that generalizes on its own; Visual adds limited but reproducible value when alignment is evaluated at the set level; and Sensory norms do not provide unique predictive power under strict out-of-fold testing.

5 Discussion

5.1 Competing Pathways of Embodiment in Chinese Color Metaphors

Our findings reveal a clear asymmetry between sensory–motor and emotional pathways in predicting cross-modal alignment for Chinese color metaphors. Sensory features such as effector exclusivity and tactile grounding showed negative correlations with alignment, and global perceptual indices were nonsignificant. Ridge regression with strict 5-fold cross-validation (with Mean/Random baselines) further indicated that sensory features achieved non-positive CV-R^2, providing no out-of-fold advantage over the mean predictor.

By contrast, emotional ratings exhibited strong and consistent positive correlations ($r \approx .45$, $p < .001$) and were the only feature group to reach consistently non-negative generalization across alignment metrics (best under trimmed-cosine, CV- $R^2 \approx 0.05$). Taken together, these results favor affective grounding over purely sensory–motor simulation for metaphorical color lexemes: affective salience affords modest but robust predictive leverage under conservative evaluation, converging with psycholinguistic evidence that affect facilitates lexical processing and memory (Barsalou, 2008; Glenberg & Kaschak, 2002; Pulvermüller, 2005). For items such as *lǜ mào* 绿帽 'betrayal', affective resonance appears to be a more reliable grounding mechanism than narrowly defined sensory-motor associations.

5.2 Generalized Experience vs. Concrete Perception

A second contrast concerns norm-based sensory ratings versus image-derived visual statistics. Whereas sensory norms largely failed to predict alignment, several low-level color statistics—notably color variability and entropy—showed moderate positive correlations (up to $r = .31$). In cross-validated prediction, visual features became informative primarily under the set-to-set Chamfer metric (CV-$R^2 \approx 0.04$), while performance with pooled-vector cosines was near zero. This pattern suggests that concrete perceptual richness in the image sets aligns better with metaphor semantics than generalized sensory norms, particularly when alignment is assessed in a many-to-many manner rather than by a single pooled vector. These observations accord with prior multimodal work linking perceptual complexity and color distributions to semantic and affective outcomes (Kiela et al., 2014; Jonauskaite et al., 2020).

6 Conclusion

We compared sensory–motor norms, emotional ratings, and image-derived visual statistics as predictors of text–image alignment for 184 Chinese color metaphors, using robust pooling and both pooled-vector and set-to-set alignment metrics. Across analyses, emotion—especially valence—was the only feature group that generalized reliably, yielding small but consistently non-negative CV-R^2 under strict 5-fold Ridge (best ≈ 0.05 with trimmed-cosine), and strong positive correlations ($r \approx .45$). Visual statistics (color variability/entropy) correlated positively with alignment and showed modest generalization primarily under Chamfer (≈ 0.04), indicating benefits when alignment is evaluated many-to-many. In contrast, sensory-motor norms did not generalize (CV- $R^2 \leq 0$) and were sometimes negatively related to alignment (e.g., effector exclusivity, tactile), suggesting that generalized sensory ratings are a poor proxy for the perceptual evidence supporting metaphor–image congruence.

Taken together, the results argue for a multidimensional embodiment in which affective salience provides a modest yet robust predictive pathway, complemented by perceptual diversity captured by visual statistics under set-to-set alignment. Methodologically, the study highlights the value of robust pooling and collection-level alignment metrics for multimodal semantics. Limitations include the absence of behavioral validation and the focus on a single metaphor family. Future work should broaden to additional metaphor domains, explore non-linear and interaction-aware models, and incorporate human judgments of lexeme-image congruence to triangulate the corpus-based findings.

References

Barsalou, L. W. (2008). Grounded cognition. Annu.Rev.Psychol., 59(1), 617-645.

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. Topics in Cognitive Science, 2(4), 716-724.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. Cognitive Neuropsychology, 33(3-4), 130-174.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. Trends in Cognitive Sciences, 15(11), 527-536.

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. Cognition, 125(3), 452-465.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. Psychonomic Bulletin & Review, 9(3), 558-565.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. Neuron, 41(2), 301-307.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. Psychonomic Bulletin & Review, 23(6), 1744-1756.

Jonauskaite, D., Parraga, C. A., Quiblier, M., & Mohr, C. (2020). Feeling blue or seeing red? Similar patterns of emotion associations with colour patches and colour terms. I-Perception, 11(1), 2041669520902484.

Kensinger, E. A. (2009). Remembering the details: Effects of emotion. Emotion Review, 1(2), 99-113.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. Journal of Experimental Psychology: General, 143(3), 1065.

Liangdong Wang, Bowen Zhang, Chengwei Wu, Hanyu Zhao, Xiaofeng Shi, Shuhao Gu, Jijie Li, Quanyue Ma, Tengfei Pan, & Guang Liu. (2024). CCI3.0-HQ: a large-scale Chinese dataset of high quality designed for pre-training large language models. CoRR, abs/2410.18505.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. Behavioral and Brain Sciences, 35(3), 121-143.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. Behavior Research Methods, 52(3), 1271-1291.

Palmer, S. E., & Schloss, K. B. (2010). An ecological valence theory of human color preference. Proceedings of the National Academy of Sciences, 107(19), 8877-8882.

Palmer, S. E., Schloss, K. B., Xu, Z., & Prado-León, L. R. (2013). Music-color associations are mediated by emotion. Proceedings of the National Academy of Sciences, 110(22), 8836-8841.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. Nature Reviews Neuroscience, 6(7), 576-582.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). Learning transferable visual models from natural language supervision. Paper presented at the International Conference on Machine Learning, 8748-8763.

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. Quarterly Journal of Experimental Psychology, 68(8), 1584-1598.

Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161.

Shutova, E., Kiela, D., & Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. Paper presented at the Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 160-170.

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. Language and Cognition, 1(2), 219-247.

Xu, X., Li, J., & Chen, H. (2022). Valence and arousal ratings for 11,310 simplified Chinese words. Behavior Research Methods, 54(1), 26-41.

Zhong, Y., Wan, M., Ahrens, K., & Huang, C. (2022). Sensorimotor norms for Chinese nouns and their relationship with orthographic and semantic variables. Language, Cognition and Neuroscience, 37(8), 1000-1022.

Language Modeling Using Entanglement Enhanced Tensor Trains

Ellis Reyes[1] **Yi-Shin Chen**[1,2]
[1]Institute of Information Systems and Applications,
[2]Department of Computer Science,
National Tsing Hua University, Hsinchu, Taiwan
ellisreyesm@gmail.com yishin@gmail.com

Abstract

Tensor Train Language Models (TTLMs) offer significant memory savings by representing text sequences as tensor networks, but naive implementations struggle with long-range dependencies and limited flexibility. We introduce a modular TTLM framework that combine local and non-local context modules to achieve scalable language modeling. Our non-local modules, inspired by entanglement in quantum information theory, enable efficient modeling of long-range interactions between hidden states. Experiments on Penn Treebank and Wikitext datasets show that our modular TTLM, including entanglement-augmented variants, outperform naive baselines. These results highlight TTLMs as a promising, memory-efficient alternatives for modern language modeling.

Keywords: Tensor Train Language Models, Entanglement-Inspired Modules

1 Introduction

Language modeling is a fundamental problem in natural language processing (Bengio et al., 2003), requiring models to capture local and global correlations across sequences of tokens. Recurrent neural networks (RNNs) (Bengio et al., 1994) and their variants (e.g., LSTMs) (Hochreiter and Schmidhuber, 1997) have been previously used, but they often fail to model long-range dependencies due to vanishing gradients. Transformers (Vaswani et al., 2017), leveraging the attention mechanism, have established new performance benchmarks and are widely used. However, their quadratic time and memory complexity in sequence length remains a bottleneck (Tay et al., 2022) (Zaheer et al., 2020).

Previous approaches of Transformer variants have been explored to tackle this issue. Linformer (Wang et al., 2020), which computes a projection of the key and value matrices to lower ranks and Longformer (Beltagy et al., 2020), which uses sparse attention patterns instead of a full dense self-attention.

Tensor Train Language Models (Su et al., 2024) have recently emerged as a theoretically memory-efficient alternative, decomposing the input sequence into a chain of low-rank tensors. These tensor networks were originally introduced in quantum many-body physics (Eisert, 2013) as efficient representations of interactions in a highly dimensional space. In contrast with Transformer variants as *Linformer* and *Longformer*, Tensor Train Language Models provide an orthogonal approach: Instead of attention sparsification, they decompose the sequential weights into a chain of low-rank tensor cores, generating a *combinatorial space* where each token is represented by a core $G^{(k)}$, constructing a global representation $G^{(1)}[i_1] \cdots G^{(d)}[i_d]$ (Equation 1). The main idea is that tensor decompositions can capture hidden correlation patterns while keeping the model scalable and interpretable, a property that has been exploited in tasks ranging from Bayesian network discovery to hierarchical clustering of real-world data (Akamatsu et al., 2025).

This property provides an analogy with language, where long-range dependencies must be captured within memory constraints. Hence, our framework adopts an interdisciplinary view. We use entanglement-inspired modules to allow TTLMs to capture dependencies, following the observation that entanglement arises from local-interactions (Eisert, 2013). Nevertheless, current naive TTLM implemen-

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 258–265
November 20-22, 2025

tations face two main limitations: (1) difficulty in modeling global, non-local dependencies, and (2) inflexibility in their core architecture, which can constrain the expressiveness of the model. To our knowledge, this is the first work that explores TTLMs beyond theory, systematically enhancing their architecture and integrating entanglement-inspired non-local modules for practical large-scale language modeling

In short, we make the following contribution:

We propose a hybrid language model architecture that combines a TTLM that captures local context via sequence processing, with causal Entanglement modules to capture non-local context (such as attention and outer product mechanisms) that allow hidden states to interact beyond local context. As a foundation for our hybrid model, we introduce standard architectural improvements such as residual connections, biases, controlled initialization and weight tying into the baseline TTLM, allowing for a stronger base model where the entanglement modules operate. Ablation studies confirm the utility of these foundational enhancements.

Experiments on the Penn Treebank and WikiText-2 datasets show that our modular TTLM, and the entanglement-augmented variants, outperform naive TTLMs. For instance, on PTB at rank 60, our best variant achieves a perplexity of 83.70 compared to 92.79 for the naive Large TTLM, a $\sim$ 9.8% reduction. A similar trend is observed on the WikiText-2, where our hybrid approach consistently outperform across most tested ranks, validating the scalability and robustness of our approach. These results suggest that TTLMs are a promising direction for scalable and memory-efficient alternatives to attention-based architectures.

2 Related Work

Beyond computational motivations, tensor networks originate from quantum many-body physics, where entanglement and entropy laws explain why high-dimensional systems can still be represented efficiently (Eisert, 2013). Recent work has begun to explore these ideas in machine learning to capture hidden correlation structures, showing tensor networks as an alternative paradigm for interpretable generative modeling (Akamatsu et al., 2025). These insights suggest that entanglement-inspired mechanisms may become a viable approach to align tensor-based architectures to capture long-range dependencies. Earlier work (Zhang et al., 2020) has explored tensor decomposition techniques as an alternative approach for language modeling to mitigate the quadratic complexity $\mathcal{O}(n^2)$ of self-attention. These methods factorize the exponential space of weight matrices into smaller tensor representations, enabling a more efficient memory scaling and inference on resource constrained hardware.

Tensor Decomposition for Language Modeling Early work (Su et al., 2024) (Zhang et al., 2020) demonstrated that tensor based representations, such as matrix product states (MPS) and tensor train (TT) approaches, can model sequences with significantly fewer parameters than Recurrent Neural Networks and Transformers. Moreover, these representations can efficiently compress LLMs components (Tomut et al., 2024) (Xu et al., 2023). The *Tensorized Transformer* (Ma et al., 2019) explore this representations using Block Term Tensor Decomposition (BTD) to achieve comparable perplexity to Transformers-XL with 2–5× fewer parameters. However, only the attention layer is compressed, the feed-forward, embedding and softmax matrices remain full size which makes the computation quadratic in complexity.

The *Tensor Train Language Model* (Su et al., 2024) introduced a fully Tensor Train architecture for language modeling as a proof-of-concept work, outperforming vanilla Recurrent Neural Networks and matching Transformer perplexity performance on limited context windows and small datasets. However, existing tensor-based approaches mainly focus on compression without providing modular architectures capable of capturing non-local dependencies. In contrast, our modular TTLM framework augmented with entanglement-inspired modules seeks to bridge the theoretical efficiency of tensor networks in quantum information with the practical requirements for language modeling. To the best of our knowledge, this entangled dynamics of local and non-local context for TLLMs has not been explored.

3 Preliminaries: Tensor Train Decomposition

The *Tensor Train* (TT) decomposition (Oseledets, 2011), factorizes a high-order tensor into a contracted sequence (or train) of low-order *cores*. Let $W \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be an order-d tensor. In TT format, each element $W(i_1, \ldots, i_d)$ is expressed as a chain of matrix products:

$$W(i_1, \ldots, i_d) = G^{(1)}[i_1]\, G^{(2)}[i_2] \;\cdots\; G^{(d)}[i_d]. \tag{1}$$

where $G^{(k)} \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$ is the k-th TT-core, $G^{(k)}[i_k] \in \mathbb{R}^{r_{k-1} \times r_k}$ corresponds the i_k-th slice along the k-th mode, and $(r_0, r_1, \ldots, r_d)$ are the TT-ranks with $r_0 = r_d = 1$. The total parameter count scales as $\sum_{k=1}^{d} r_{k-1} n_k r_k$, which grows linearly with d, in contrast to the exponential $\prod_k n_k$ parameters of a dense tensor.

3.1 Tensor Train Language Models

Tensor Train Language Models use the principles of TT decomposition introduced in Section 3 to generate a parameter-efficient approach for sequence modeling. Following a standard RNN update rule $h_t = f(x_t, h_{t-1})$, where $h_t \in \mathbb{R}^R$ is the hidden state of dimension R (TT-rank or rank), x_t is the input token at time t, and f is the cell function implemented using the tensor contractions.

The input token x_t is first mapped to an embedding vector E'_t. For example, in small variants like *TTLM-Tiny* (Su et al., 2024) The hidden state h_{t-1} is transformed by an input-to-hidden learnable weight matrix W_{ih}, and the result computes a matrix-vector product with E'_t, followed by an activation function:

$$h_t = (h_{t-1} W_{ih} + b_{ih}) \cdot E'_t \tag{2}$$

Larger variants like *TTLM-Large* (Su et al., 2024) first process the input embedding E_t by an additional hidden-to-hidden weight matrix W_{hh} before reshaping it into the embedding E'_t:

$$h_t = (h_{t-1} W_{ih} + b_{ih}) \cdot \mathrm{reshape}(E_t W_{hh}) \tag{3}$$

This allows for a more detailed input dependent state transition.

4 Methodology

We propose a Tensor Train Language Model that converges (i) a recurrent TTLM for local context and (ii) non-local *entanglement* modules that enable hidden states interaction to capture long-range dependencies.

Figure 1 illustrates the overall architecture: The TTLM processes the input sequence $(x_1, \ldots, x_N)$ recurrently (left), where the state at each time h_t depends on the previous state h_{t-1} and the current input (x_t). The resulting sequence of hidden states, $h_1, \ldots, h_N$, where N represents the total length of the input sequence, is then aggregated by causal entanglement modules (right) to capture non-local dependencies before making a final prediction.

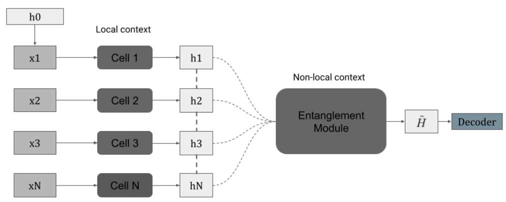


Figure 1: Modular TTLM with Entanglement.

Our key assumption here is that by entangle the local and non-local modules the expressiveness of the model increases. Our intuition is by allowing the hidden states to interact globally with non-local past context, we expect to capture long-range dependencies by the connection of the current state and its history. Moreover, the modularity could enable ablation studies to compare different behaviors among the entanglement variants to analyze outcomes for specific types of data and tasks.

4.1 Modular TTLM (Local Context)

Our hybrid architecture introduces an enhanced Tensor Train Language Model based on the TTLM-Large implementation from Section 3.1 and equation 3. The hidden state $h_t \in \mathbb{R}^R$ is updated from h_{t-1} and the current embedding $E_t \in \mathbb{R}^{R^2}$ as follows:

$$
\begin{aligned}
v_t &= (h_{t-1}W_{ih} + b_{ih}) \\
E'_t &= \text{Reshape}(E_t W_{hh} + b_{hh}) \\
h_t &= \text{Activation}\big(v_t \cdot E'_t + \alpha \cdot h_{t-1} + b_{\text{cell}}\big)
\end{aligned} \tag{4}
$$

In Equation 4, the previous state is transformed by a linear computation (W_{ih}, b_{ih}) where W_{ih} is a learnable weight matrix that connects input and hidden states, this matrix determines which past memory slices are relevant for the current position. The transition matrix E'_t is computed by reshaping the current embedding E_t by another linear computation (W_{hh}, b_{hh}). Here, W_{hh} is a learnable weight matrix that determines how the current input should influence the update. The reshaping flatten the transition matrix E'_t into $R \times R$. Conceptually, E'_t represents a flattened TT-core slice $G^{(k)}[i_k]$ from the decomposition chain seen in Eq. 1.

With this approach, we separate the relevant past memory slices into the transformed state v_t and the current state influence into the transition matrix E'_t. Hence the core interaction involves the matrix-vector product $v_t \cdot E'_t$. The flow of the gradient is improved with a residual connection to preserve information by adding the previous state back scaled by a hyperparameter α. In this implementation the biases (b_{ih}, b_{hh}) modify the inputs in the core tensor contraction computation and the bias b_{cell} is used as an extra learnable parameter to shift the overall output independently of the internal linear transformations before it passes through an activation function such as tanh. This activation introduces non-linearity, though omitting it (using a linear activation) aligns with baseline configurations (Su et al., 2024).

We employ weight tying for efficiency to share the weights between E_t and the output projection layer. All learnable parameters within the cell (the entire computational unit) $[W_{ih}, b_{ih}, W_{hh}, b_{hh}, b_{\text{cell}}]$ are initialized using a scaled uniform distribution $[-0.1/R, 0.1/R]$ for smooth initialization and stability, especially at higher ranks.

The Modular TTLM processes the input sequence $X = (x_1, \ldots, x_N)$ sequentially. The resulting sequence of hidden states

$$H = [h_1, h_2, \ldots, h_N] \in \mathbb{R}^{N \times R}$$

captures the local context and serves as the input to the causal Entanglement Blocks described in the next section.

4.2 Entanglement Modules (Non-local Context)

While the Modular TTLM introduced in Section 4.1 processes local dependencies, because of the recurrent nature potential information from the distant positions could be lost. To tackle this limitation and capture non-local dependencies, we introduce the Entanglement modules. From information theory perspective, Shannon entropy measures the expected information gain before observing the outcome (Baez, 2024), hence our intuition for the entanglement modules is to reduce this uncertainty by allowing hidden previous states to share information. These modules are designed to be causal, hence the computation of the output feature for time t only depends on the input hidden states up to that time step.

Chunked Low-Rank Attention. Given the hidden states $H \in \mathbb{R}^{N \times R}$, we divide each sequence of length N into $M = \lceil N/C \rceil$ non-overlapping chunks of size C. This summaries are linearly projected to obtain the keys and values. The chunked attention update is given by:

$$\tilde{H} = H \; + \; \text{softmax}\Big(\frac{QK_{\text{proj}}^\top}{\sqrt{R}} \; + \; \mathbf{C}\Big)V_{\text{proj}}, \tag{5}$$

where $\mathbf{C}$ is the causal mask. For each hidden state at time t belonging to chunk j all entries with $m \geq j$ are masked, hence each query can attend only to summaries of earlier chunks conserving causality.

This implementation reduces the complexity from $O(N^2R)$ to $O(NMR)$. Like Linformer (Wang et al., 2020), the method is low-rank in sequence length, but instead of learned projection matrices, it uses a pooling compression where each chunk summary is the mean of the TTLM hidden states within that window. These pooled representations provide a compact and non-local summary of the past context.

Causal Hadamard Pooling. Given hidden states $H \in \mathbb{R}^{N \times R}$, we use an Exponential Moving Average (EMA) vector $\mathbf{e}_t$ for each sequence in the batch:

$$\mathbf{e}_t = \alpha\, \mathbf{e}_{t-1} + (1-\alpha)\, h_t, \qquad \alpha = \sigma(\lambda),$$

where λ is a learnable scalar shared across layers and σ denotes the sigmoid function. In this implementation, the EMA is a summary of all past hidden states, hence each new update depends only on the previous average and the current input. Each output state is updated by a Hadamard interaction with its EMA given by:

$$\tilde{h}_t = h_t + g_{\text{ent}}\left(h_t \odot \tfrac{1}{\sqrt{R}}\mathbf{e}_t\right), \tag{6}$$

where g_{ent} is a learnable gating parameter and $1/\sqrt{R}$ normalizes the interaction. A light feed-forward network finalizes the representation to get the output $\tilde{H}$. This computation scales linearly $O(NR)$, since the EMA is computed once per step and reused. Conceptually, the module performs a causal outer-product interaction between the current state and a low-rank summary of the past, allowing each position to integrate non-local information.

4.3 Prediction and Training

We obtain an enriched hidden sequence $\tilde{H} = [\tilde{h}_1; \ldots; \tilde{h}_N]$ which integrates both local and non-local context as described in Sections 4.2 and 4.1. Each state is projected to the vocabulary space with a linear decoder:

$$\ell_t = W_{\text{dec}}\, \tilde{h}_t + b_{\text{dec}}, \qquad p(y_t \mid x_{<t}) = \text{softmax}(\ell_t). \tag{7}$$

Here, $W_{\text{dec}} \in \mathbb{R}^{V \times R}$ and $b_{\text{dec}} \in \mathbb{R}^{V}$ are the parameters of the output projection, where V is the vocabulary size. The linear transformation maps the hidden state $\tilde{h}_t$ to the logits ℓ_t, which are then converted into a probability distribution over the vocabulary using a softmax function. The model is trained using standard cross-entropy loss over all time steps:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{t=1}^{N} \log p(y_t \mid x_{<t}). \tag{8}$$

This training objective optimizes the conditional probability of the next token.

5 Experimental Setup

We implemented our models in PyTorch (Paszke et al., 2019) using the PyTorch Lightning framework (Falcon and The PyTorch Lightning team, 2019). We evaluated our approach on a classic word-level and open domain language modeling task using (1) the Penn Treebank (PTB) dataset (Marcus et al., 1994), which contains 929k tokens for training, 73k for validation, and 82k for testing, with a vocabulary size of 10k unique words. PTB is suitable to evaluate compact language models with short context length. And (2) The WikiText-2 (WT2) (Merity et al., 2016), which contains 2.1M tokens for training, 218k for validation, and 246k for testing, with a vocabulary size of about 33k words. WT2 provides a larger and more natural vocabulary than PTB which makes it suitable to evalute language models for longer dependencies. The main evaluation metric is perplexity (PPL) on the test set, where lower perplexity indicates better language modeling performance. Specifically, we compare the following models:

1. TTLM Baselines variants: TTLM-small and TTLM-large (Su et al., 2024).
2. Modular TTLM (Ours): Modular TTLM-small and Modular TTLM-large.
3. Modular TTLM + Entanglement (Ours): Chunked Low-Rank Attention and Hadamard Pooling.
4. Transformer Baseline: Transformer with L blocks, multi head self-attention, pre-norm, learned positional embeddings. Configured to match the parameter count of the corresponding TTLM variant (Vaswani et al., 2017).

All models were trained under identical optimization settings for fairness. To compare TTLM baselines and our Modular TTLM, we used a learning rate of 0.001, dropout rate of 0.25, batch size of 32 and Adam optimizer. For our Modular TTLM variants, additional architectural hyperparameters were set to $\alpha_{\text{res}} = 0.5$ and use_tanh = False.

For the Modular TTLM + Entanglement aditional parameters were used such as AdamW optimizer with weight decay 0.2, dropout rate of 0.5, a 1000 steps warm up, a chunk size of 10 and entanglement_gate_init at 0.0.

To ensure a fair comparison, the Transformer baseline is sized to match the total parameter count of the corresponding Modular TTLM-Large variant. We choose the embedding dimension, number of layers, Heads and feed-forward width such that the amount of parameters closely matching the TTLM run at rank R. For example, at rank 60 we choose a $(1024, 7, 8, 2048)$ transformer run. For the matching at rank R the embedding dimension and number of layers were modified, Heads and feed-fordward width stayed constant at 8 and 2048.

All experiments were performed for up to 50 epochs with early stopping, gradient clipping at 0.25, a batch size of 32, sequence length (bptt) of 35, embedding size of 400 and a fixed random seed of 42 was used for all runs to ensure reproducibility. All experiments were conducted on one NVIDIA A100-SXM4-80GB GPU.

5.1 Scaling Law Evaluation

To analyse the scaling properties of the proposed Modular TTLM variants, we train each model with increasing TT-rank values ($R \in \{5, 10, 15, 20\}$) and report the corresponding test perplexity for a fair comparison with the baseline. This experiment highlights how the expressiveness of the Modular TTLM improves with rank, in analogy to scaling laws observed in large scale language models.

6 Results

6.1 Scaling with TT-Rank

We first study the effect of increasing TT-rank on language modeling performance. Table 1 reports test perplexity on PTB for the baselines variants (Su et al., 2024): Large *TTLM-L*, Small *TTLM-S* and our variants: Modular Large *M-TTLM-L*, Modular Small *M-TTLM-S*. Both modular models consistently outperform their corresponding baselines across severall ranks, with the largest improvements observed at higher ranks. This shows that the proposed architectural enhancements improve the expresiveness capabilities of the model.

Table 2 shows the results on WikiText-2, where the same trend is observed: increasing rank lowers perplexity, Modular TTLM outperform the baselines. This shows the robustness of the models indicating that modeling global interactions is helpful for the larger WT2 vocabulary.

Rank	TTLM-L	TTLM-S	M-TTLM-L	M-TTLM-S
5	156.0	161.7	151.7	169.8
10	119.9	127.5	116.4	127.8
15	106.7	117.1	103.3	116.2
20	102.1	111.2	**96.7**	**109.2**

Table 1: Perplexity on PTB test set across TT-ranks.

Rank	TTLM-L	TTLM-S	M-TTLM-L	M-TTLM-S
5	129.1	132.9	127.8	140.0
10	100.4	109.4	100.0	109.3
15	89.4	100.9	88.1	100.8
20	85.5	96.7	**82.0**	**95.5**

Table 2: Perplexity on WikiText-2 test set across TT-ranks.

6.2 Effect of Entanglement Modules at Higher Ranks

We next compare our Modular TTLM-Large baseline (*M-TTLM-L*) against its entanglement-augmented variants: *M-TTLM-A* (chunked low-rank attention) and *M-TTLM-H* (Hadamard pooling) for higher ranks ($R \in \{40, 60, 70\}$). The scaling experiments in Section 6.1 showed that performance generally improves with rank, motivating our expectation that entanglement modules could further enhance expressiveness by capturing non-local context. Tables 3 and 4 show results on PTB and WikiText-2.

Both entanglement modules yield modest yet consistent improvements over our Modular TTLLM-Large, with *M-TTLM-A* performing best on PTB and WikiText-2. All our variants outperform the baseline TTLM-Large and Transformer on PTB. We attribute the poor improvement on WikiText-2 at higher ranks to the short context window used in our experiments ($bptt = 35$).

WikiText-2 contains longer sequences and a more diverse vocabulary than PTB, which likely makes it more sensitive to context windows. This could also explains why Transformer models perform better on this dataset. While improvements are modest, the results hint that causal non-local interactions might offer a way toward increasing the expressiveness of recurrent language modeling.

Rank	M-TTLM-L	M-TTLM-A	M-TTLM-H	Trans	TTLM-L
40	87.2	86.1	85.9	87.8	95.7
60	85.1	83.7	84.0	83.6	92.8
70	85.2	84.4	84.6	86.2	93.5

Table 3: Perplexity on PTB across higher ranks for Modular TTLM-Large (M-TTLM-L) and entanglement variants.

Rank	M-TTLM-L	M-TTLM-A	M-TTLM-H	Trans	TTLM-L
40	75.1	74.0	74.2	70.2	77.8
60	74.2	73.5	73.6	73.7	79.7
70	74.9	74.5	73.6	69.4	81.5

Table 4: Perplexity on WikiText-2 across higher ranks for Modular TTLM-Large (M-TTLM-L) and entanglement variants.

6.3 Efficiency Analysis

To evaluate computational requirements, we measure training wall time and GPU memory usage at TT-rank 60 for PTB and WikiText-2 (Table 5). We observe that the introduction of entanglement modules does not significantly increase training or inference GPU memory consumption which suggests that the proposed modules can be integrated with minimal computational cost while maintaining scalability as introduced in the formulation of the entanglement modules in the method section 4.

Our Rank 60 model demonstrates performance on PTB (Table 3) and WikiText-2 (Table 4) that is comparable to the Transformer baseline. Table 5 highlights the efficiency advantages, showing lower inference memory usage and similar or lower training memory usage compared to the Transformer. It is important to note that while parameter counts are comparable, the Transformer hyperparameters (embedding dimension, layers, heads) were chosen to match our TTLM model size. Performance may vary with different Transformer configurations, as Transformer performance is sensitive to these architectural choices.

6.4 Scaling Context Length on WikiText-2

To explore how our model performs and scales on a larger and diverse dataset like the WIkitext-2, we investigated the output of increasing the context length available to the model during training. We run separate instances of our Modular TTLM, with rank $R = 60$ a sequence lengths *bptt* of 128 and

Model	Params	Train (MB)	Infer (MB)
PTB (rank = 60)			
M-TTLM-L	49.3M	1047.7	281.7
M-TTLM-A	49.3M	1048.2	281.7
M-TTLM-H	49.3M	1048.0	281.7
Transformer	51.2M	1423.5	888.9
WikiText-2 (rank = 60)			
M-TTLM-L	117.4M	2600.4	696.7
M-TTLM-A	117.4M	2600.9	696.7
M-TTLM-H	117.4M	2600.7	696.7
Transformer	118.0M	2414.9	2051.8

Table 5: GPU memory usage for Modular TTLM-Large and its entanglement variants at TT-rank 60.

Model	Params	PPL	Train (MB)	Infer (MB)
WikiText-2 (bptt=128)				
M-TTLM-L	117.3M	67.20	2887.9	1353.4
M-TTLM-A	117.3M	66.42	2905.6	1353.3
M-TTLM-H	117.3M	66.58	2897.3	1353.4
WikiText-2 (bptt=256)				
M-TTLM-L	117.5M	66.49	4429.6	2258.3
M-TTLM-A	117.5M	65.38	4591.3	2258.9
M-TTLM-H	117.5M	65.95	4448.7	2258.9

Table 6: WikiText-2 perplexity, number of parameters, and GPU memory usage as a function of context length.

256. For these runs, the chunk size for the low-rank attention entanglement was set to 16 and 32, respectively to investigates if there are increments on memory usage. Table 6 presents the peak GPU memory usage during training and inference as we increase the sequence length. When doubling the context length from 128 to 256, the peak memory usage for all models increases by a factor of approximately $1.5 - 1.7$. This observed scaling is substantially less than the quadratic increase expected from standard self-attention mechanisms and is consistent with the theoretically linear ($O(L)$) in our hybrid architecture for both M-TTLM-A and M-TTLM-H and the M-TTLM-L. The trend observed between $bptt = 128$ and $bptt = 256$ provides strong empirical evidence supporting the linear memory complexity claim for our proposed models.

7 Conclusion

Our preliminary results indicate that our Hybrid approach for language modeling improves the expressiveness capabilities of TTLM based architectures improving perplexity on the Penn Treebank and WikiText-2 Dataset. These gains come with little to no increase in GPU memory or wall time, consistent with our theoretical complexity analysis. This suggest that TTLMs could incorporate non-local context while retaining their quasi-linear memory scaling advantages over non recurrent models.

For future work, we plan to extend our evaluation to larger and diverse datasets. Additionally, we aim to explore a different range of entropy-based modules to study the implications of entanglement principles in modeling language structure.

References

Katsuya O Akamatsu, Kenji Harada, Tsuyoshi Okubo, and Naoki Kawashima. 2025. Plastic tensor networks for interpretable generative modeling. *arXiv preprint arXiv:2504.06722*.

John C Baez. 2024. What is entropy? *arXiv preprint arXiv:2409.09232*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Jens Eisert. 2013. Entanglement and tensor network states. *arXiv preprint arXiv:1308.3318*.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

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

Xueyun Ma, Peng Zhang, Sheng Zhang, Nan Duan, Yue Hou, Ming Zhou, and Daxin Song. 2019. A tensorized transformer for language modeling. In *Advances in Neural Information Processing Systems*, volume 32.

Mitch Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Ivan V Oseledets. 2011. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Zhenning Su, Yunlong Zhou, Fengwei Mo, and Jakob G. Simonsen. 2024. Language modeling using tensor trains. *arXiv preprint arXiv:2405.04590*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28.

Andrei Tomut, Seyed Saeed Jahromi, Abhinaba Sarkar, Ugur Kurt, Suraj Singh, Faraz Ishtiaq, Carlos Muñoz, Prateek S. Bajaj, Ahmed Elborady, Alberto Del Bimbo, et al. 2024. Compactifai: extreme compression of large language models using quantum-inspired tensor networks. *arXiv preprint arXiv:2401.14109*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Mengyuan Xu, Yulan Lin Xu, and Danilo P. Mandic. 2023. Tensorgpt: Efficient compression of large language models based on tensor-train decomposition. *arXiv preprint arXiv:2307.00526*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Shuai Zhang, Peng Zhang, Xindian Ma, Junqiu Wei, Ningning Wang, and Qun Liu. 2020. Tensorcoder: Dimension-wise attention via tensor representation for natural language modeling. *arXiv preprint arXiv:2008.01547*.

Multimodal Fake News Detection Combining Social Network Features with Images and Text

Lawrence Y. H. Low, Yen-Tsang Wu, Yan-Hong Liu, Jenq-Haur Wang
Department of Computer Science and Information Engineering
National Taipei University of Technology
Web Information Retrieval Lab
t113999402@ntut.org.tw t107599005@ntut.edu.tw
t112598044@ntut.org.tw jhwang@ntut.edu.tw

Abstract

The rapid development of social networks, coupled with the prevalence of Generative AI (GAI) in our society today, has led to a sharp increase in fake tweets and fake news on social media platforms. These fake media led to more in-depth research on fake news detection. At present, there are two mainstream methods used in detecting fake news, namely content-based fake news detection and propagation / network-based fake news detection. Early content-based detection method inputs an article's content and uses a similarity algorithm to identify fake news. This method improved by using single-modality features such as images and text as input features. However, existing research shows that single-modality features alone cannot identify fake news efficiently. The most recent method then fuses multimodal features such as images and text, as features to be input into the model for classification purposes. The second propagation / network-based fake news detection method creates graphs or decision trees through social networks, treating them as features to be input into the model for classification purposes. In this study, we propose a multimodal fake news detection framework that combines these two mainstream methods. This framework not only uses images and text as input features but also combines social metadata features such as comments. The framework extracts these comments and builds them into a tree structure to obtain its features. Furthermore, we also propose different feature fusion methods which can achieve better results compared with the existing methods. Finally, we conducted ablation experiments and proved that each module is required to contribute to the framework's overall performance. This clearly demonstrated the effectiveness of our proposed approach.

Keywords: Fake news detection, Multimodal fusion, Multimodal learning, Social media

1 Introduction

With the development of the Internet, social media has replaced most traditional media such as newspapers and magazines. Although these network developments provide convenience for users (Khattar et al., 2019), due to the increasingly rapid speed of information dissemination on social networks and the development of Generative AI (GAI), fake news is generated in increasing quantities and its spread is also transforming significantly. Therefore, how to quickly and effectively classify fake news is an important research topic.

Over the years, online social media content has evolved from plain text to multimodal content that combines pictures and text, and in some cases, videos, sounds, and text (He et al., 2021; Pinnaparaju et al., 2021). Early research on fake news detection mainly used single-modality classification, but studies have found that single-modality can no longer efficiently detect fake news. Therefore, current research topics are moving towards using multimodality features to detect fake news. Multimodality refers to the combination of multiple types of modalities, including but not limited to pictures, text, and sound. Chen et al. (2022), and Singhal et al. (2019) reported that past multimodal fake news detection tends to solve the fake news problem by considering additional subtasks such as event discriminators and finding cross-modal correlations. Fake news detection relies heavily on subtasks. Without subtasks, the detection results

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 266–276
November 20-22, 2025

will be significantly reduced. It was found that cross-modal content can now provide additional supplementary features for fake news detection, but these studies mainly focus on the integration of cross-modal content. Previous studies did not take into account the differences between content in different modalities, resulting in poor model performance. The Modular Co-Attention Network (MCAN) method is inspired by the way humans read news with pictures and text (Wu et al., 2021). They proposed a multimodal joint attention network and found that the interdependence between multimodal features can achieve better detection results.

However, different modalities may express the same thing at certain times. In this case, adding multimodal fusion features will create noise and affect the performance of the classification task. On the other hand, when the unimodal detection performance is subpar, multimodal fusion features may be added to increase the feature input for better model performance. Therefore, researchers should be aware of the impact of the modifications between different modalities on the model. Other than the usage of unimodal features, a timely addition of multimodal features can obtain better classification results (Qian et al., 2021; Zhou et al., 2020; Wang et al., 2018), therefore most recent research studies focus on how to fuse different modalities and understand the consistency between different modalities to achieve better accuracy.

Currently, almost all research on multimodal fake news detection uses images and text. However, with the rapid development of GAI, the generation of fake news has become faster and more compelling, therefore using only images and text as input is no longer sufficient to accurately detect fake news. More diverse features must be considered to assist fake news detection, but relatively few studies have incorporated social metadata features. In addition to user information such as “retweets”, “likes”, and “number of friends”, common social metadata features also use shared comments or articles to establish a propagation path structure of a graph or tree structure (Rahimi et al., 2024; Li et al., 2020). However, these existing methods rarely combine the features of the two methods.

In order to solve the above-mentioned problems, this study refers to the architecture MMFN, Multi-grained Multi-modal Fusion Network (Zhou et al., 2023) and proposes a novel framework, Multi-Modal Title Comment (MMTC) combining social metadata features and multimodal fusion of text and images. In addition to the integration of unimodal features and multimodal fusion features of text and images, our proposed method adds social metadata features with comment tree structure as input to achieve more accurate fake news detection. The MMTC framework includes:

1) Multimodal fusion module: The module obtains single-modal feature input through pre-trained BERT (Devlin et al., 2019) and Swin Transformer v2 (Liu et al., 2022). The pre-trained model CLIP (Radford et al., 2021) is used to extract semantic information between different modalities to solve the problem of semantic inconsistency between the different modalities.
2) Title and comment module: The similarity between the text and the image summary is used as a weight, multiplied by the features of the comment tree structure and subsequently concatenated with the image summary features to evaluate whether the social background feature is important based on the relevance between the image and the text.

We summarize our key contributions in this paper as follows:

- We propose a novel framework, MMTC that uses social network comments, pictures, and text as features, taking into account both the details and overall aspects of the news.
- We demonstrate the effectiveness of our framework by comparing with existing baseline methods using well-known datasets. MMTC outperforms the existing fake news detection methods.
- We perform ablation tests to verify the effects of the various modules in our framework are effective.

2 Related Works

The purpose of fake news detection is to distinguish the authenticity of news based on the relevant information of the news released on social media platforms. This information may include text content, image content, comments, communication structure and other user characteristics. Related research can be divided into two categories based on the data. The first category is based on article content. The features of this method usually use the content of the article, such as pictures, text, and in some cases, news videos. The second category is based on social background. This method uses the information about the user in the news as features, such as numerical features such as "retweets", "likes", and "friends". It also uses graphs or tree structures to transform the user's information into the dissemination structure of the article in order to use it as a social interaction feature.

2.1 Content-based fake news detection

In recent years, fake news has been spreading frequently on social media platforms. According to previous studies (Liu and Wu, 2020; Shu et al., 2017; Rubin et al., 2016), it is crucial to detect fake news. Research on fake news detection can be divided into two broad categories: Based on (i) news content and (ii) social context. The method based on news content can be sub-divided into unimodal fake news detection and multimodal fake news detection.

Unimodal Fake News Detection Previous studies on fake news detection have mostly focused on single modality, with a large portion of them using text content analysis (Nan et al., 2021; Ajao et al., 2019) and image content analysis (Jin et al., 2017). The amount of existing information makes traditional manual detection more difficult, and fake news detection models based on Machine Learning is used to mitigate this limitation. Ma et al. (2016) proposed a method that uses Recurrent Neural Networks (RNN) to learn features. The results show that models based on Deep Learning are more effective. Kaliyar et al. (2021) reported several methods of fake news detection such as Features-based approaches, Knowledge-based approaches, Learning-based approaches, and proposes a BERT-based method that only uses text data as input training. Xue et al. (2020) proposed a Multi-Vision Fusion Neural Network (MVFNN) for the detection of fake news pictures, combining the pixel domain, frequency domain and tampering detection features of the image.

Multimodal Fake News Detection Although fake news can be effectively identified by simply using text or pictures. Online social platforms include rich multimodal information such as picture, text, video (Zhang et al., 2019) and uses existing post datasets to achieve multimodal fake news detection by extracting visual emotion features, text emotion, behavioral responses, and metadata (Leung et al., 2023). Ying et al. (2023) reported that the consistency between cross-modalities and the features of different modalities affecting model decisions are still unresolved. The authors proposed a method of extracting features from different perspectives of text, image patterns, and image semantics, and using the representation of each image to approximately predict the authenticity of the news. This multimodal representation can predict the consistency across the different modalities, thereby obtaining accurate fake news detection. Qi et al. (2023) focused on short videos to detect fake news. Their model added social metadata features such as "comment", "user information" and used a Cross-Model Transformer to learn the relationship between different modalities. These added social metadata features are in addition to the video content features such as pictures, text, and images. Palani et al. (2022) used images and text, together with CapsNet and BERT as input models to extract features in order to combine the features for fake news classification.

2.2 Fake news detection based on social context

The interaction between social media allows news to have a variety of social activities. For example, after the news is released, users can share, discuss and analyze it with their friends. These constitute the social interaction of news, which includes not only the authenticity of news reports, but also users' comments on the news and their emotions towards the news. Usually, social metadata features are obtained using structure-based methods. User information can be obtained from social media, and unstructured data can be combined into structured data such as graphs and tree-structures. Graph based methods have achieved remarkable results because they can closely simulate the social interaction and the process of spreading online news.

Related work on graph methods includes Qian et al. (2016), who propose a novel Multi-modal Multi-view Topic Opinion Mining (MMTOM) model for social event analysis in multiple collection sources. MMTOM can effectively combine multimodal and multi-view attributes in a unified and principled manner for social event modeling. It not only discovers multimodal common topics from all collections and summarize the similarities and differences of these collections on each specific topic, but also automatically mine multi-perspective opinions on the learned topic in different collections. Gong et al. (2023) presented a systematic survey of graph-based fake news detection research and Deep Learning-based techniques. We further discuss the challenges and unsolved issues in graph-based fake news detection and identification as well as the future research directions. Zhang et al. (2021) proposed a model based on Graph Attention Networks to extract information from user interactions. In the communication graph, nodes represent user text content and edges represent response interactions. The authors implemented an attention mechanism to decide the edge weights between pairs of nodes.

Regarding tree-structed research methods, Ni et al. (2021) aim to solve the problem of fake news detection in real-world scenarios. The authors developed a new Neural Network-based model to detect fake news and provide explanations on social media. Only the source short tweet, and its retweets are provided as features, however user comments are omitted. A Multi-task Attention Tree Neural Network (MATNN), proposed by Bai et al. (2023) jointly classify stance and detect the authenticity of rumors. The authors designed a structural representation, which converts irregular rumor conversation trees into Regular rumor Conversation Trees (RC-Trees). When extracting features, the authors use the tree's word attention mechanism to extract local structures for stance analysis. Zhang et al. (2023) organized claim posts in a cycle as a temporary event tree, extracts event elements, and converts them into bipartite graphs of temporary event trees in terms of posts and authors, namely, author tree and post tree. We propose a novel rumor detection model with a hierarchical representation on a bipartite temporal event tree.

3 Model architecture and methods

We propose a multimodal framework, Multi-Modal Title Comment (MMTC) that combines social network features to improve multimodal models that only use content or only use social metadata. MMTC consists of two modules: Multi-Modal Block (MMB) and Title-Comment Block (TCB). MMB (section 3.1) contains two types of inputs. Unimodal feature input, which includes the unimodal input of images and texts in the architecture to represent the overall representation of texts and images. Multimodal fusion input, which represents the detailed input of the consistency of images and texts in the architecture. TCB (section 3.2) includes a similarity weight calculation and a social background feature input. The image title similarity weight, which evaluates the importance of the comment features of the article by calculating the similarity between the image summary and the title text, and the comment tree structure feature, which calculates the comment features related to the article by establishing a tree structure. The final portion of MMTC is a fake news classifier (section 3.3). The overall architectural design of MMTC is as shown in Figure 1. MMTC has two major blocks: Multi-Modal and Title-Comment. The feature strings from these two blocks are sent to the classifier for training and classification.

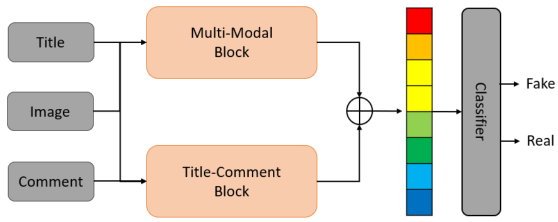


Figure 1. Overall Multi-Modal Title Comment (MMTC) framework.

3.1 Multi-Modal Block (MMB)

The MMB module is divided into two parts: Unimodal feature input and Multimodal feature input. The unimodal features are the title text of the tweet and the image in the tweet. The pre-trained model is used to extract the overall semantic representation of the single modality in the model. In the multimodal feature, the pre-trained multimodal model is used to find the correspondence between words and images, and subsequently, the multimodal detailed semantic representation is obtained through the designed Multiple Feature Fusion network. The Multi-Modal Block (MMB) architecture is as shown in Figure 2.

In MMB, the unimodal inputs of text and images will use a pre-trained BERT and Swin Transformer v2 respectively, whereas the multimodal input will use the pre-trained CLIP model. The unimodal features will be added to the text or image output generated by the multimodal through average pooling, and finally the feature output will be obtained through the projection head. The multimodal part will be input into a multimodal fusion block, Multiple Feature Fusion to obtain feature output.

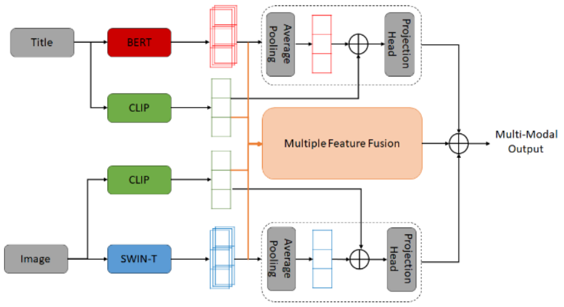


Figure 2. Multi-Modal Block (MMB) architecture.

Unimodal Feature Input The unimodal text features here use the pre-trained BERT to extract text features. BERT is a Transformer based model developed by Google that uses a large amount of unlabeled text in an unsupervised manner. It is one of the most commonly used language models. Therefore, we use BERT here to extract text features in the text content. The text of the title will be input into BERT and output to the last layer of the hidden layer, and average pooling will be used to compress the features to obtain the overall text semantic representation as FT.

The unimodal image features are obtained using a pre-trained Swin Transformer v2, an upgraded version of Swin Transformer (Liu et al., 2021). The upgraded version is a hierarchical Visual Transformer architecture designed for efficient processing of high-resolution images and various downstream tasks such as classification, detection, and segmentation. In this paper, the image is pre-processed and the image size is converted to 224 × 224. The image is input into the last layer of the output hidden layer of Swin Transformer v2. Similar to text processing, average pooling is used to compress the features to obtain the overall image semantic representation as FI.

Multimodal Feature Input The purpose of multimodal feature input is to find the corresponding relationship between different modalities to obtain features. In this paper, the CLIP model is used to obtain multimodal features. CLIP is a novel multimodal model proposed by OpenAI in 2021. Its pre-training task allows the model to predict from scratch on a dataset containing 400 million image-text pairs, from which the model learns the caption to obtain image feature representation.

Our proposed method use CLIP to extract two different embeddings, CLIP text embedding and CLIP image embedding. These two embeddings represent the overall semantic vector of the sentence and the overall semantic vector of the image, respectively. Since single-modal feature inputs FT and FI cannot directly realize information interaction, a multi-modal fusion method, Multiple Feature Fusion Block (MFFB) is designed here to obtain information interaction between single modality and multimodality, as shown in Figure 3.

The MFFB uses the Co-attention Transformer method (Lu et al., 2019) to experiment with cross-modal information complementarity using the features extracted from BERT and Swin Transformer v2. MFFB uses the co-attention for single modal input to obtain the focus of attention of the other modality, and then connects the outputs between the two modalities and multiplies the weighted similarity of the two inputs of CLIP to obtain the multimodal fusion feature.

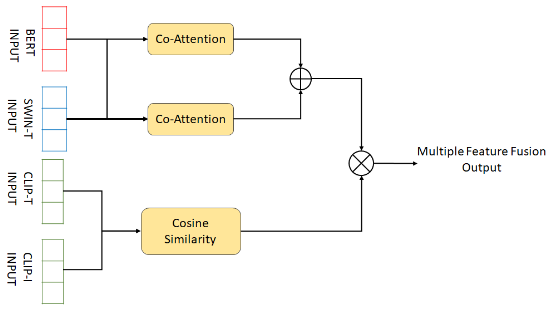


Figure 3. Multiple Feature Fusion Block (MFFB).

The Co-attention Transformer architecture is as shown in Figure 4.

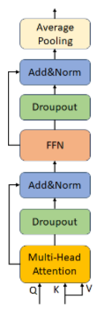


Figure 4. Co-attention Transformer architecture.

We treat the two modalities F1 as Query and F2 as Key and Value and input them into the Co-attention Transformer. Q is the input of one modality, K and V are the input of the other modality. Using this method, the input modality can obtain the attention focus of the other modality.

In this way, we can obtain two different representations. One is the focus of attention in the image guided by the text, read as $FT2I$, and the other is the focus of attention in the text guided by the image, read as $FI2T$. The two features of the previous CLIP are matched through Cosine Similarity to find out whether the two are matched, which is converted into a weight, read as WC. Finally, $FI2T$ and $FT2I$ are concatenated and multiplied by WC to obtain the final output $FMMF$ of MFFB, as shown in Eq. (1):

$$F_{MMF} = (F_{I2T} \oplus F_{T2I})\, W_C \qquad (1)$$

The feature output FMM of Multi-modal Block consists of three features concatenated together as shown in Eq. (2):

$$FMM = F_T \oplus F_I \oplus F_{MMF} \qquad (2)$$

3.2 Title-Comment Block (TCB)

The Title-Comment Block (TCB) uses the similarity between image summary and title to discover the degree of relevance between image and text. It then weights against the comment feature composed of the comment tree, and finally concatenates with the image summary to obtain the feature output. The TCB consists of two blocks. The first block is the similarity calculation between the article title and the text summary. The second block is the comment tree feature established by the article comments. It is weighted by the similarity weight to distinguish which comments are relatively important. Finally, the image summary and comment features are concatenated as the output features of the TCB. The TCB architecture is as shown in Figure 5.

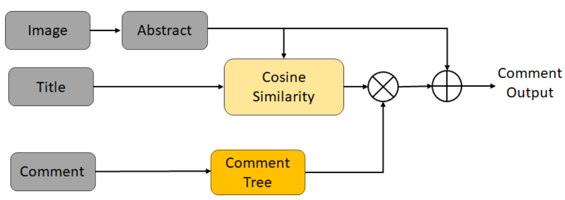


Figure 5. Co-attention Transformer architecture.

Image and Title Similarity Weight In order to find the appropriate social metadata features to reduce noise, similarity calculation is used as weighted weight to calculate the text summary of the article title and the picture, respectively. The text summary of the picture is obtained through the pre-trained multimodal model BLIP (Li et al., 2022). BLIP is a multimodal learning architecture proposed by the Salesforce team. Its purpose is to enhance multimodal performance by aligning images with text. BLIP can perform image question answering, image description generation and multimodal classification tasks. In the paper, we used BLIP to obtain the text summary of the picture and perform cosine similarity calculation with the article title to obtain the image-text semantic similarity weight WIT of the article.

Comment Tree-structured Features In the propagation/network-based methods based on social metadata, comments are often used as one of the features. Some methods treat comments as text and input them into the model together with other text features (Kirchknopf et al., 2021), while others create a tree structure for comments (Ma et al., 2018), which allows people to reply to comments posted by others on social media platforms.

In our proposed method, the article title is taken as the root of the tree structure, and the reply comments are used to establish the leaf nodes of the tree. According to the reply to each comment, a structure tree around the reply-article-title will be established by using a recursive method to calculate from bottom to top. The comment tree-structured features are as shown in Figure 6. In the example, comment 3 replies to comment 1, whereas comment 1 and comment 2 reply to the article title. The features of the parent node are used in a bottom-up manner using a multi-head attention or self-attention, and then the comment feature output of the article is calculated in a cascade manner.

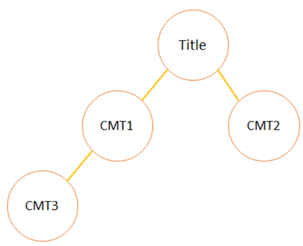


Figure 6. Comment Tree-structured Features.

Since it is impossible for every article on the social platform to have a reply, there are two situations: (i) articles with no comments and (ii) articles with at least one comment. For articles with no comments, our proposed method sets the part without comments to a 256-dimensional tensor vector with all zeros, and uses self-attention to calculate the comment features represented by the tree. Whereas for articles with at least one comment, the parent node of each layer represents each child node and performs multi-head attention calculation with the parent node. Finally, the output of the horizontal child node is concatenated to represent the feature output of the parent node. The comment feature FC represented by the final tree, that is, the article, is calculated through recursion. The feature output of Title-Comment Block is FTC, as shown in Eq. (3):

$$FTC = WIT \times FC + FIA \quad (3)$$

3.3 Fake News Classifier

The fake news classifier in this paper is input from two modules in the architecture, namely the multimodal feature input FMM of the Multiple Feature Fusion Block and the comment social metadata feature FTC of the Title-Comment Block. We concatenate the two feature inputs as shown in Eq. (4):

$$F = FMM \oplus FTC \quad (4)$$

Then we feed the final features into the classifier, as shown in Eq. (5):

$$y = classifier(F) \quad (5)$$

The classifier consists of three fully connected layers and a ReLU activation function. F is the concatenation of all modules. The classification method uses Focal Loss (FL) to classify the prediction result as true or false, as shown in Eq. (6):

$$FL(y) = -\alpha(1 - y)\gamma log(y) \quad (6)$$

α is the balance parameter between positive and negative samples, γ is the penalty for easily classified samples, and y is the probability of predicting the correct category.

4 Experimental Setup and Results

In this section, we present the various parameter settings of our experiments, explain the baseline models and our experimental results.

4.1 Experiment Setup

Datasets Fakeddit is a multimodal fake news dataset released in 2020 (Nakamura et al., 2020). The data is collected from Reddit which contains text, pictures, raw data and comments. The data were collected between March 2008 to October 2019. The data have multiple labels, namely 2-way, 3-way and 6-way. In our experiments, 2-way labels are used for model training, and the features used are "clean title", "title id", "image", "comment", and "comment id" from the dataset.

Preprocessing Because our proposed model simultaneously uses text, images, and comments, we first filtered the dataset to include only articles containing both text and images. We applied a second filter based on the number of comments and replies. However, since not all articles receive replies, we selected those with a reply count between 0 and 5 comments for this study. For the image summaries, we employed the BLIP model to generate captions ranging from 5 to 20 words. After filtering, the dataset size is about 160,000. We randomly selected 50,000 data as our experimental dataset, in which the ratio of true to false news is set to 1:1.

Evaluation Metrics We use accuracy as the evaluation criterion for the binary classification of fake news. Additionally, precision, recall, and F1 score are also used as supplementary evaluation criteria.

Experimental Details In our experiments, we set a dimension of 512. For co-attention, the embedding dimension is 256, and FFN hidden dimension is 512, with 8 multi-heads and dropout is set to 0.1. The text embedding dimension of BERT is set to 256, and uses "bert-base-uncased" model for English data. The maximum input text length is 512. For Swin Transformer v2, we use "microsoft/swinv2-tiny-patch4window8-256" for image features, and the input image size is set to 224*224. The CLIP model uses "openai/clip-vit-base-patch32" with both features set to 256.

The unimodal features of text and images are added together after the output features of CLIP and then reduced to 256 using a linear layer. The features output by multiple feature fusion are also reduced to 256 dimensions. Finally, MMB concatenates text features, image features, and multimodal features and reduces the dimension to

256. In TCB, BERT is also used for text features. The image summary and comment features are concatenated, and then the dimension is reduced to 256. Finally, the features of the two modules are concatenated, and the dimension is reduced to 256.

During the training phase, we scaled the number of classifier layers to 64, 16, and 2, respectively. The batch size is 1. The optimizer uses AdamW, the Learning Rate is set to 1e-6, and the weight drop is set to 1e-1. The model is trained for up to 20 epochs, and early stopping is used to prevent overfitting. For *FL* Loss Function setting, we set alpha to 1 and gamma to 0.25.

4.2 Baseline models

We compare our experimental results with existing method proposed by Uppada et al. (2022). Their method uses Fakeddit dataset with images, text and emotional features to train the model. However, our proposed model does not include emotional features as part of the social metadata; instead, it focuses on images and textual content. We use cross-entropy as our Loss function.

4.3 Experimental results

We present the model comparison between our method and other existing baseline methods. The experimental results showed that our model, MMTC with added social network features achieve better results than models that only uses image and text as input. The results of our method, MMTC and other baseline methods is as shown in Table 1, best results are marked in bold.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F1	Precision	Recall	F1
(BERT+Dense) +Xception	0.826	0.809	0.859	0.833	0.846	0.793	0.819
MMTC	**0.861**	**0.878**	**0.843**	**0.86**	**0.846**	**0.88**	**0.863**

Table 2: Baseline model comparison.

We present the ablation test results that we performed on each input feature of our model, MMTC. From the ablation test results, we observed that, simply by using the multimodal model CLIP, it has the best accuracy among all the modules. However, it is still slightly lower than the accuracy of our proposed framework. An additional finding from the experimental result is the accuracy of using only comments as features is much lower than that of directly obtaining features from the article. Instead, by combining the entire module, the framework can achieve better results. This indicates that the comment features and image summary features can help our model to achieve better classification results. The ablation test results for each modality are as shown in Table 2, best results are marked in bold.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F1	Precision	Recall	F1
Text	0.812	0.869	0.74	0.799	0.769	0.886	0.824
Image	0.777	0.792	0.757	0.774	0.762	0.797	0.779
Multiple	0.856	0.873	0.836	0.854	0.839	0.876	0.857
Image Abstract	0.652	0.651	0.668	0.66	0.652	0.635	0.644
Comment	0.487	0.444	0.059	0.104	0.491	0.925	0.641
MMTC	**0.861**	**0.878**	**0.843**	**0.86**	**0.846**	**0.88**	**0.863**

Table 1: Ablation test results.

5 Conclusion and future work

In this study, we propose a multimodal model that incorporates social network features for fake news detection. Based on previous research, we believe that content-based methods cannot efficiently detect fake news, hence we add a comment tree structure based on social metadata features to assist the detection task. We also designed a different feature fusion method, relying on the similarity of image and text associations to weight single-modal features. Our model also achieves better results than existing methods using the same dataset, and ablation tests on each feature are performed to prove that our inputs are necessary.

For future work, we plan to address the limitations of our proposed model, MMTC through continued research. We aim to use additional social metadata features, such as emotional stance and short videos, to enhance the diversity of the model. To improve the model, we plan to use the full Fakeddit dataset, conduct ablation studies on the effects of comments, and investigate gradient accumulation techniques. For better model evaluation, we intend to compare with recent SOTA models and more up-to-date datasets. With the current advancement of large language models, generative AI, and explainable AI with enhanced reasoning capabilities, we aim to improve our fake news detection model by extending our research in this direction.

Acknowledgments

The authors would like to thank the support from the National Science and Technology Council, Taiwan under the grant numbers: NSTC 114-2221-E-027-068, NSTC 114-2634-F-027-001-MBK.

References

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/ abs/2103.00020

Armin Kirchknopf, Djordje Slijepčević, and Matthias Zeppelzauer. 2021. Multimodal Detection of Information Disorder from Social Media. In *2021 International Conference on Content-Based Multimedia Indexing* (*CBMI*). 1–4. doi:10.1109/ CBMI50038.2021.9461898

Balasubramanian Palani, Sivasankar Elango, and Vignesh Viswanathan K. 2022. CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT. *Multimedia Tools Appl.* 81, 4 (Feb. 2022), 5587–5620. doi:10.1007/s11042-021-11782-3

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference* (San Francisco, CA, USA) (*WWW'19*). Association for Computing Machinery, New York, NY, USA, 2915–2921. doi:10.1145/3308558. 3313552

Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multimodal Knowledge-aware Event Memory Network for Social Media Rumor Detection. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) (*MM'19*). Association for Computing Machinery, New York, NY, USA, 1942–1951. doi:10.1145/3343031.3350850

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Curran Associates Inc., Red Hook, NY, USA.

Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. Exploiting Microblog Conversation Structures to Detect Rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online),5420–5429. doi:10.18653/v1/2020.coling-main.473

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 1980–1989. doi:10.18653/v1/P18-1184

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) (*IJCAI'16*). AAAI Press, 3818–3824.

Johnathan Leung, Dinusha Vatsalan, and Nalin Arachchilage and. 2023. Feature analysis of fake news: improving fake news detection in social media. *Journal of Cyber Security Technology* 7, 4 (2023),224–241. doi:10.1080/23742917.2023.2237206 arXiv:https://doi.org/10.1080/23742917.2023.2237206

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086 [cs.CV] https://arxiv.org/abs/2201.12086

Junxiao Xue, Yabo Wang, Shuning Xu, Lei Shi, Lin Wei, and Huawei Song. 2020. MVFNN: Multi-Vision Fusion Neural Network for Fake News Picture Detection. In *Computer Animation and Social Agents*, Feng Tian, Xiaosong Yang, Daniel Thalmann, Weiwei Xu, Jian Jun Zhang, Nadia Magnenat Thalmann, and Jian Chang (Eds.). Springer International Publishing, Cham, 112–119.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 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 (Eds.). European Language Resources Association, Marseille, France, 6149–6157. https://aclanthology.org/2020.lrec-1.755/

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (Sept. 2017), 22–36. doi:10.1145/3137597.3137600

Marzieh Rahimi and Mehdy Roayaei. 2024. A Multi-View Rumor Detection Framework Using Dynamic Propagation Structure, Interaction Network, and Content. *IEEE Transactions on Signal and Information Processing over Networks* 10 (2024), 48–58. doi:10.1109/TSIPN.2024.3352267

Na Bai, Fanrong Meng, Xiaobin Rui, and Zhixiao Wang. 2023. A multi-task attention tree neural net for stance classification and rumor veracity detection. *Applied Intelligence* 53, 9 (May 2023), 10715–10725. doi:10.1007/s10489-022-038335

Nikhil Pinnaparaju, Manish Gupta, and Vasudeva Varma. 2021. T3N: Harnessing Text and Temporal Tree Network for Rumor Detection on Twitter. In *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021*, Virtual Event, May 11–14, 2021, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, 686–700. doi:10.1007/978-3-030-75762-5_54

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment Aware Fake News Detection on Online Social Networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2507–2511. doi:10.1109/ICASSP.2019.8683170

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. FakeSV: a multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence* (*AAAI'23/IAAI'23/EAAI'23*). AAAI Press, Article 1620, 9 pages. doi:10.1609/aaai.v37i12.26689

Qi Zhang, Yayi Yang, Chongyang Shi, An Lao, Liang Hu, Shoujin Wang, and Usman Naseem. 2023. Rumor Detection with Hierarchical Representation on Bipartite Adhoc Event Trees. arXiv:2304.13895 [cs.SI] https://arxiv.org/abs/2304.13895

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23).* AAAI Press, Article 601, 9 pages. doi:10.1609/aaai.v37i4.25670

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-domain Fake News Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) (*CIKM'21*). Association for Computing Machinery, New York, NY, USA, 3343–3347. doi:10.1145/3459637.3482139

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools Appl.* 80, 8 (March 2021), 11765–11788. doi:10.1007/s11042-02010183-2

Santosh Kumar Uppada, Parth Patel, and Sivaselvan B. 2022. An image and text-based multimodal model for detecting fake news in OSN's. *J. Intell. Inf. Syst.* 61, 2 (Nov. 2022), 367–393. doi:10.1007/s10844-022-00764-y

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical Multi-modal Contextual Attention Network for Fake News Detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) (*SIGIR'21*). Association for Computing Machinery, New York, NY, USA, 153–162. doi:10.1145/3404835.3462871

Shengsheng Qian, Tianzhu Zhang, and Changsheng Xu. 2016. Multi-modal Multi-view Topic-opinion Mining for Social Event Analysis. In *Proceedings of the 24th ACM International Conference on Multimedia* (Amsterdam, The Netherlands) (*MM'16*). Association for Computing Machinery, New York, NY, USA, 2–11. doi:10.1145/2964284.2964294

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. SpotFake: A Multi-modal Framework for Fake News Detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data* (*BigMM*). 39–47. doi:10.1109/BigMM.2019.00-44

Shiwen Ni, Jiawen Li, and Hung-Yu Kao. 2021. MVAN: Multi-View Attention Networks for Fake News Detection on Social Media. *IEEE Access* 9 (2021),106907–106917. doi:10.1109/ACCESS.2021.3100245

Shuzhi Gong, Richard O. Sinnott, Jianzhong Qi, and Cecile Paris. 2023. Fake News Detection Through Graph-based Neural Networks: A Survey. arXiv:2307.12639[cs.SI]https://arxiv.org/abs/2307.12639

Victoria L. Rubin, Yimin Chen, and Nadia K. Conroy. 2016. Deception detection for news: Three types of fakes. In *Proceedings of the Association for Information Science and Technology* 52, 1 (Feb. 2016), 1–4. doi:10.1002/pra2.2015.145052010083

Xinpeng Zhang, Shuzhi Gong, and Richard O. Sinnott. 2021. Social Media Rumour Detection Through Graph Attention Networks. In *2021 IEEE Asia Pacific Conference on Computer Science and Data Engineering*(*CSDE*).1–6. doi:10.1109/CSDE53843.2021.9718466

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multimodal Fake News Detection. In *Advances in Knowledge Discovery and Data Mining*, Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan (Eds.). Springer International Publishing, Cham, 354–367.

Yang Liu and Yi-Fang Brook Wu. 2020. FNED: A Deep Network for Fake News Early Detection on Social Media. *ACM Trans. Inf. Syst.* 38, 3, Article 25 (May 2020), 33 pages. doi:10.1145/3386253

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2560–2569. doi:10.18653/v1/2021.findings-acl.226

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multi-modal Fake News Detection on Social Media via Multi-grained Information Fusion. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval* (Thessaloniki, Greece) (*ICMR'23*). Association for Computing Machinery, New York, NY, USA, 343–352. doi:10.1145/3591106.3592271

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) (*KDD'18*). Association for Computing Machinery, New York, NY, USA, 849–857. doi:10.1145/3219819.3219903

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW'22)*. Association for Computing Machinery, New York, NY, USA, 2897–2905. doi:10.1145/3485447.3511968

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022. Swin Transformer V2: Scaling Up Capacity and Resolution. arXiv:2111.09883 [cs.CV] https://arxiv. org/abs/2111.09883

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 [cs.CV] https://arxiv.org/abs/2103.14030

Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor Detection on Social Media with Event Augmentations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) (*SIGIR'21*). Association for Computing Machinery, New York, NY, USA, 2020–2024. doi:10.1145/3404835.3463001

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Transactions on Multimedia* 19, 3 (2017), 598–608. doi:10.1109/TMM.2016.2617078

基於語音指令的中文 ASR 錯誤校正系統設計與實現
Speech-Driven Editing System for Chinese ASR Errors

丁仕杰 **Sji-Jie Ding**, 張嘉惠 **Chia-Hui Chang**, 簡資烜 **Zi-Xuan Jian**
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering, National Central University
shijied675@gmail.com, chiahui@g.ncu.edu.tw, 114522129@cc.ncu.edu.tw

摘要

儘管人工智慧技術近年進步顯著，但 ASR 系統仍難以應對現實中由發音和同音詞引起的錯誤。爲解決此問題，本研究提出一種基於語音指令的校正方法。系統由三個模組組成：輸入分類器、命令分類器和校正標註器。爲了支援訓練和評估，我們透過 TTS 和 ASR 流程模擬 ASR 錯誤，以模擬潛在的錯誤，並基於漢字形聲字語言特徵及大型語言模型產生校正命令。實驗顯示基於兩種校正指令的混合模型, 整體校正準確率超過 80% 且性能穩定。與手動校正相比，雖然較慢但仍有一定競爭力，顯示其實際部署的可行性。

Abstract

Despite recent advances in AI, ASR systems still struggle with real-world errors from pronunciation and homophones. To solve this issue, we propose a verbal-command-based correction system that enables users to utter natural-language instructions to refine recognition outputs with minimal effort. The system consists of three modules: an input classifier, a command classifier, and a correction labeler. To support training and evaluation, we simulate ASR errors via TTS and ASR pipelines to simulate the potential errors, followed by verbal correction commands issued based on linguistic features or LLMs. Experiments show that the overall system achieves over 80% correction accuracy and delivers stable performance. Compared to manual correction, this system also demonstrates highly competitive correction speed, which sufficiently indicates its feasibility for practical deployment.

關鍵字：語音辨識、錯誤修正、語音指令、自動修正模組、中文自然語言處理

Keywords: Automatic Speech Recognition, Voice Command, Error Correction

1 緒論

隨著自動語音辨識（ASR）技術的發展，語音輸入已逐漸融入日常生活，從智慧家電、語音助理到訊息傳遞，都能看見它的身影。

然而，儘管中文 ASR 在公開測試集的字元錯誤率（CER）已降至 3.05% (Xu et al., 2025)，但因個人發音差異及大量同音異字，實際錯誤率遠高於預期。例如，用户説「這個程式很棒」，可能被誤識別爲「這個城市很棒」。

爲此，許多相關研究提出了自動錯誤修正的方法。早期方法透過額外特徵（如聲學資訊 (Zhang et al., 2021)、多模態輸入 (Jiang et al., 2024a)）增強端到端模型。爲緩解序列到序列模型常見的過度修正問題（即引入新錯誤），有研究提出如 PGCC (Dong et al., 2024) 等框架。近年來，大型語言模型（LLM）已被應用於此任務，透過融合拼音等方式提升表現 (Li et al., 2025; Wei et al., 2024)。

與此相關的拼寫錯誤校正（SEC）任務，也發展出預測編輯操作（如 LASERTAGGER (Malmi et al., 2019)）或整合語言聲學線索（如 D^2C (Jiang et al., 2024b), ReLM (Liu et al., 2024)）的方法。近期 LLM 雖可透過提示詞技術 (Yang et al., 2023; Li et al., 2024) 應用於此，但仍需手動驗證。

綜上所述，雖然自動修正技術已取得顯著進展，但當修正失敗時，使用者仍需依賴鍵盤輸入，對於不擅長打字或雙手不便的情境並不友善。爲此，我們提出一套基於語音指令的 ASR 錯誤修正系統，允許使用者直接透過口説指令修改辨識錯誤，以減少鍵盤操作並提升使用體驗。

同時爲了模擬眞實場景，我們利用文句轉語音（TTS）與 ASR 建構資料集以生成潛在錯誤，並設計了符合中文使用者習慣的修正指令。本研究主要貢獻如下：

- 提出以口説指令修正 ASR 錯誤的系統，減少手動編輯，提升互動效率。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 277–285
November 20-22, 2025 ©2025 Association for Computational Linguistics

- 系統支援多種指令類型（新增、修改、刪除），並整合分類、解析、標註到自動編輯的統一流程。
- 利用 TTS 與 ASR 建構了多樣化的語音指令與場景資料，模擬辨識錯誤，產出高品質資料集以強化系統。

2 系統架構與資料集

本節將說明本研究所提出之語音辨識錯誤修正系統的整體架構與資料處理流程。該系統主要由三個核心模組組成：輸入分類器（Input Classifier）、指令分類器（Command Classifier）以及指令標註器（Command Labeler），分別負責識別輸入意圖、判斷修正類型，並產生對應的修正指令。在資料準備方面，本研究以現有文字資料集爲基礎，透過文字轉語音（Text-to-Speech, TTS）與語音辨識（Automatic Speech Recognition, ASR）技術模擬語音辨識錯誤，進一步設計多樣化的修正敘述，並依據三個模組的功能需求整理成對應的訓練資料集。以下將依序介紹各模組功能與資料處理方式。

2.1 系統架構

圖 1 爲我們系統的整體架構。我們將整個流程分成了三個模組，分別是輸入分類器、指令分類器和指令標註器。透過這三個模組，就能依照使用者提供的指令來提取各項資訊，修正原先的文字。

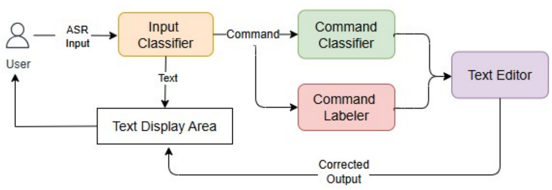


圖 1: 系統流程圖

2.1.1 輸入分類器

輸入分類器的作用爲分辨當前的語音輸入內容屬於單純的文字訊息，或是用來修改前文的指令。我們以 0 和 1 兩個標籤來表示文字和指令兩個不同的類別。當模型判斷該內容爲指令時，就會執行後續的指令分類器和指令標註器，提取指令修改需要的資訊。模型選擇的部分我們使用 BERT 的中文模型，如圖 2 所示。

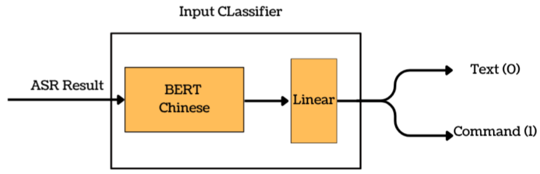


圖 2: 輸入分類器架構

2.1.2 指令分類器

我們將輸入分類器判斷爲指令的內容，當作輸入提供給指令分類器，讓模型判斷該指令屬於替換 (Replace)、新增 (Insert) 或是刪除 (Delete) 這三種類型的哪一個類別，其對應的標籤分別爲 0、1、2。和輸入分類器相同，模型我們一樣選擇使用 BERT 的中文模型，如圖 3 所示。

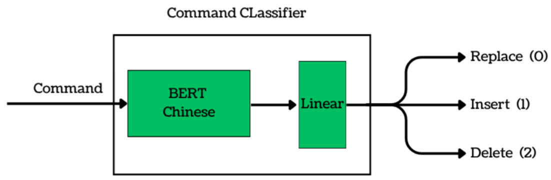


圖 3: 指令分類器架構

2.1.3 指令標註器

最後一個模組，我們將要修改的文字 (Text) 和指令 (Command) 結合，中間以 [SEP] 區隔，一起當作輸入提供給模型。模型會先對輸入進行斷詞 (tokenize)，接著判斷每一個 token 屬於”O”、”B-Modify” 或是”B-Filling” 的哪一個類別。”B-Modify” 會出現在要修改的文字中的某個位置，”B-Modify” 則會出現在指令中的某個位置，其餘無關的部分會以”O”來表示，整體架構如圖 4 所示，與另外兩個模組不同的地方在於，模型我們選擇了 BERT 的多語言模型。

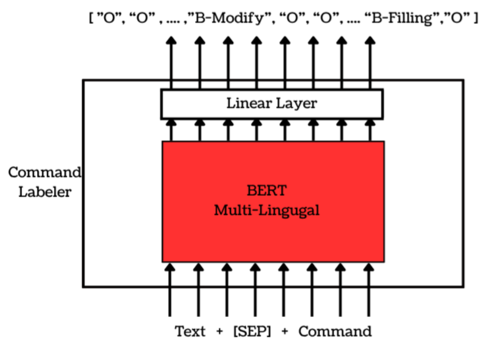


圖 4: 指令標註器架構

經過這三個模組之後，我們會得到指令類型、修改範圍和填入字這三個重要的元素，依

據不同的指令類型，我們使用不同的應對方法進行修改。舉例來說，如果是替換的指令，我們就會將修改範圍的內容和填入字直接進行替換；如果是刪除，則就單純刪除修改範圍的內容；最後針對新增的指令，我們會在修改範圍的前一個位置插入標註的填入字。對於整體架構來說，這三個模組都是不可或缺的一部分。

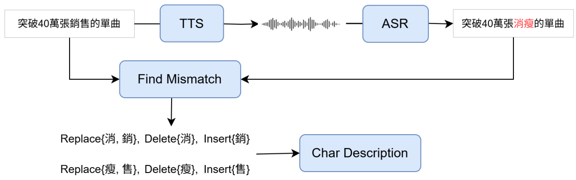


圖 5: 錯誤資料 & 刪除、新增錯誤產生流程

2.2 資料集準備

對於上述三個模組的訓練資料，我們選擇了兩個現有資料集來當作基礎，分別是 SIGHAN-2015[1] 和 zh-tw-wikipedia[2]。SIGHAN 是針對中文拼字檢查任務的資料集，內容包含錯誤句子、正確句子以及修正步驟，因此被廣泛使用。zh-tw-wikipedia 則包含了截止至 2023 年 5 月，中文維基百科 2,533,212 篇條目的文字內容，我們先將每篇文章的內容進行斷句，並過濾掉中文以外語言的內容。為了模擬語音辨識產生的輸出，我們將這些資料透過雅婷文字轉語音[3]的文字轉語音功能，將文字轉換成音訊檔案。選擇該服務的原因在於其合成語音提供了較接近於台灣本土的口音，更加貼合現實情境。得音訊檔案後，我們再使用 Google 的 Speech-to-Text AI [4] 的語音轉文字功能，將音訊檔轉換回文字，以此來模擬使用者利用語音辨識得到的文字內容，如圖5上半部所示。

對於每筆原始句子和經過 TTS 和 ASR 過後的結果，我們計算兩者之間的 Levenshtein distance (Levenshtein et al., 1966)，以此來得知我們需要對錯誤句子進行多少次 Replace, Delete 和 Insert 的動作將其還原成正確的句子。然而，分析過後，我們大多數的修正動作都是 Replace，因為語音辨識的錯誤較少發生多字或是少字的情況，這會導致我們後續產生指令資料的比例不平均。為了解決這個問題，我們將一部份 Replace 指令拆分成 Delete 和 Insert 兩個動作。舉例來說，原先的修正方式是將” 冰” 這個字替換成” 濱”，拆分過後變成先刪除” 冰” 字，再新增” 濱” 這個字，等同於直接進行替換的動作，解決資料不平衡的問題，如圖 5 下半部所示。

2.2.1 指令資料生成

我們希望每則指令能更貼近真實使用情境。以中文字為例，當需要修改某個字時，使用者可能會透過部件組合或常見詞語來描述，例如：「弓長張」、「耳東陳」、「祝福的祝」、「辛苦的苦」等。為了模擬這樣的敘述風格，我們針對不同資料集設計了不同的指令產生方式。

在 SIGHAN-2015 資料集中，我們使用 GPT-4o 並下達提示語，要求模型以常用詞語來描述修改內容，例如前述的「祝福的祝」，藉此產生完整的指令格式（詳見附錄6.1）。

至於 zh-tw-wikipedia 資料，我們依據以下優先順序，產生原字與替換字的敘述方式：

1. 部首不同但同音的字（例如：「人部的他」vs. 「女部的她」）
2. 可拆解為具體部件的字（例如：「弓長張」）
3. 使用常用詞彙進行描述

為實現上述規則，我們從「CNS11643 中文標準交換碼全字庫」[5]彙整了常用字的注音、部首與部件資訊，以判斷是否滿足條件一和二。至於常用詞彙描述，我們不依賴大型語言模型，而是參考標準的常用字詞列表；若單一字元對應多個候選詞彙，則隨機擇一使用。表 1 展示了由這兩種方法生成的指令範例。

Command	Type
把人部的佛改成弓部的弗	Replace
請在民主的民前面新增製造的造	Insert
刪除自心息	Delete

表 1: 指令範例

2.3 各模組資料類型

我們從上述資料中分別建立了三個不同的資料集，用於訓練我們系統的三個模組，分別是輸入分類器、指令分類器和指令標註器。

2.3.1 輸入類型資料集

我們將每筆 Wrong text 和其對應的指令分別標示為 0 和 1，如表 2 所示，代表文字和指令兩個類別，即可用於輸入分類器模組的訓練，辨識使用者當前說的內容屬於其中哪一項。

[1]http://sighan.cs.uchicago.edu/
[2]https://huggingface.co/datasets/zetavg/zh-tw-wikipedia
[3]https://tts.yating.tw/
[4]https://cloud.google.com/speech-to-text?hl=zh-TW
[5]https://www.cns11643.gov.tw/

Input	Type	Label
我知道，你很久以前找工作很幸苦。	Text	0
請把幸福的幸改成辛苦的辛。	Command	1

表 2: 輸入類別範例

2.3.2 指令類型資料集

指令則分成三個類別，分別是替換、新增和刪除，對應的標籤爲 0, 1, 2，如表 3。此資料集將使用於指令分類器的訓練，因爲我們需要確認指令的類型爲何，才能在最後修改的步驟根據其類型來做出對應的修正。

Command	Type	Label
請把發財的發改成爬山的爬。	Replace	0
請在民主的民前面新增製造的造。	Insert	1
刪除週末的週。	Delete	2

表 3: 指令類別範例

2.3.3 指令標註器資料集

當我們得到生成錯誤句子和對應的指令資料後，就能得知錯誤句子中的修改範圍，還有指令中的填入字爲何。因此我們將錯誤句子和指令串接在一起，中間使用 [SEP] 符號作爲分隔。接著我們使用 google-bert/bert-base-multilingual-cased [6]的 Tokenizer 將整合的句子分割成 token，將修改範圍的位置標示爲 B-Modify，填入字的位置標示爲 B-Filling，其他部分則標示爲 O。選擇該 Tokenizer 的原因在於後續訓練我們皆會使用 BERT 系列的模型，其中 multilingual 的模型對於數字以及英文的斷詞有較好的能力。在指令標註器的訓練過程，我們將使用 Sequence2Tag 的模型，幫助我們成功標記出這兩個關鍵的元素。以下爲資料範例，Input 前半段的內容，「你的回答我受到了，謝謝。」，爲需要修改的句子；後半段的句子，「刪除受到的受」，則爲修正指令，兩者中間加入 [SEP] 區分。Output Label 則表示輸入進行斷詞後，經由模型輸出的標記結果，B-Modify 對應的位置爲「你的回答我受到了，謝謝。」中的「受」字，B-Filling 則對應指令中「刪除受到的受」最後的「受」字。

- **Input:** 你的回答我受到了，謝謝。[SEP] 刪除受到的受
- **Output Label:** ['O', 'O', 'O', 'O', 'O', 'B-Modify', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-Filling']

[6]https://huggingface.co/google-bert/bert-base-multilingual-cased

3 實驗

3.1 實驗設置

在實驗中，我們依據 Section 2.2 所述的方法，爲每個模組建立所需的訓練資料，並針對 SIGHAN 與 zh-tw-wikipedia 兩個資料集分別訓練模型。如表 4 所示，我們亦採用兩者混合的 Mix dataset，以增強模型對不同指令敘述方式的適應能力。

本研究選用 BERT（Bidirectional Encoder Representations from Transformers）(Devlin et al., 2019) 系列模型進行訓練。BERT 採用基於 Transformer 的雙向編碼器架構，透過遮蔽語言建模（MLM）與下一句預測（NSP）進行預訓練，具備優秀的語意理解能力，並能靈活應用於分類與序列標記等多項 NLP 任務。在本研究中，BERT 能有效處理錯誤位置辨識與指令類型分類兩種任務。

我們使用的模型爲 google-bert/bert-base-chinese [7] 以及 google-bert/bert-base-multilingual-cased，因爲兩者已在大規模中文語料上預訓練，很適合處理繁體中文文本。

Task	SIGHAN-15		zh-tw-wikipedia		Mix dataset	
	Train	Test	Train	Test	Train	Test
Input Type Classifier	4264	474	5202	578	9826	1052
Command Classifier	1990	222	3321	369	5311	591
Command Labeler	2194	244	4510	501	6704	745

表 4: 各模組訓練&測試資料數量

3.2 輸入分類器

在輸入分類器的實驗中，我們使用基於 BERT 的中文預訓練模型進行訓練。資料總量如表 4 第一列所示。實驗結果如表 5 所示。模型名稱以訓練資料集命名，例如 Model-SIGHAN 表示使用 SIGHAN-15 資料集訓練的模型。

模型在不同資料集上皆能準確地判斷輸入內容爲純文字輸入或語音指令。儘管在混合資料集上的 F1 分數略降至 0.99，整體表現仍維持高準確度。此分類器的判斷能力對於後續系統的正確運作至關重要。

Model	Model-SIGHAN			Model-Wiki			Model-Mix		
	P	R	F1	P	R	F1	P	R	F1
BERT Chinese	1.0	1.0	1.0	1.0	1.0	1.0	0.99	0.99	0.99

表 5: 輸入分類器校能 (Performance of the Input Classifier)

[7]https://huggingface.co/google-bert/bert-base-chinese

3.3 指令分類器

在指令分類器的實驗中，我們需進一步判斷指令類型爲「替換」、「新增」或「刪除」，以便後續根據類型執行對應處理。資料總量如表 4 第二列所示。本實驗同樣使用基於 BERT 的中文預訓練模型進行訓練，實驗結果如表 6 所示。

與輸入分類器結果相同，模型在三種類型的指令分類任務中皆達到極高準確度，於三種資料集（SIGHAN、Wiki、Mix）上之 Precision、Recall 與 F1-score 均爲 1.0，顯示模型能穩定且正確地辨識不同類型的指令。值得注意的是，混合資料集（Model-Mix）包含不同的指令敘述方式，但模型在此仍維持完美表現，反映其在異質資料上的良好泛化能力。

這樣的結果顯示，模型不僅能學習清楚區分指令類型的語言特徵，也具備跨語域資料的適應能力，有助於提升後續修正的準確度。

Model	Model-SIGHAN			Model-Wiki			Model-Mix		
	P	R	F1	P	R	F1	P	R	F1
BERT Chinese	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

表 6: 指令分類器效能

3.4 指令標註器

指令標註器的任務需同時精準標示兩類位置資訊：原文中需修改的字詞（B-Modify），以及指令中欲替換的新字詞（B-Filling）。資料總量如表 4 第三列所示。本任務使用 mBERT（Multilingual BERT）模型進行訓練，主因是需標註的句子中可能包含非中文字詞，例如英數字或特殊符號，相較於前兩項使用的中文 BERT 模型，多語言模型在此具更佳的語言適應性。

實驗結果如表 7 所示，模型在三組資料集上皆展現穩定表現，F1 值介於 0.92 至 0.94 之間。SIGHAN 資料集上 F1 爲 0.92，略低於 Wiki 資料集的 0.94，推測可能與 SIGHAN 中錯字形式更爲多樣，標註任務較具挑戰有關；而混合資料集 Model-Mix 亦能維持 F1 score 0.93 的水準，顯示模型具一定泛化能力。

Model	Model-SIGHAN			Model-Wiki			Model-Mix		
	P	R	F1	P	R	F1	P	R	F1
mBERT	0.91	0.92	0.92	0.95	0.92	0.94	0.93	0.93	0.93

表 7: 指令標註器效能

3.5 系統整體效能

我們遵循圖1所示之流程，並使用「指令標註器」(Command Labeler) 的測試集來評估此框架。如表8所示，Model-SIGHAN 與 Model-Wiki 兩個模型在其領域內皆修正了超過 80% 的句子，但在交叉測試中則表現出較差的泛化能力。錯誤分析指出，大部分的失敗案例源於標註問題，以及對不熟悉的指令描述存在理解上的困難。相比之下，Model-Mix 在不同資料集上均取得了穩定的準確率，這表明混合資料的訓練方式能有效提升模型的泛化能力。這些發現凸顯了爲改善未來系統，收集更多樣化指令描述的重要性。

Model	SIGHAN (244)		Wiki (501)	
	Match	Acc	Match	Acc
Model-SIGHAN	198	0.81	191	0.39
Model-Wiki	145	0.59	420	0.84
Model-Mix	200	0.82	419	0.84

表 8: 不同模型在兩個資料集上的修正準確率與匹配數（Match / Acc），資料集後面的數字爲測試資料總比數。

3.6 語音指令辨識錯誤的影響

在資料集建構時，我們針對文字輸入進行錯誤模擬，建立了目前的 SIGHAN-15 與 zh-tw-wikipedia 兩個訓練資料集。然而，在實際應用中，使用者以語音下達指令的時候，可能因語音辨識錯誤而影響系統判斷。爲此，我們將上述兩個資料集中 Command Labeler 的指令，同樣透過文字轉語音（TTS）與語音轉文字（ASR）處理，產生模擬語音輸入錯誤的資料。我們將這些資料當作新的測試集，並依照上一小節的實驗設計，進一步分析在加入可能含有語音辨識錯誤的指令（Error Cmd）後，對整體流程準確率的影響。

實驗結果如表9 所示。可觀察到，在處理語音辨識錯誤的情況下，原始 Model-Mix 模型效能明顯下降；在 SIGHAN 測試集的準確率僅爲 0.64，而在 Wiki 測試集更低至 0.02，顯示語音辨識錯誤對部件／部首描述的指令影響尤爲顯著。

Model	SIGHAN		Wiki	
	Match / Acc	ΔAcc	Match / Acc	ΔAcc
Model-Mix	156 / 0.64	–	12 / 0.02	–
Model-Mix + NoisyCmd	151 / 0.62	-0.02	222 / 0.44	+0.42

表 9: 不同模型架構於語音錯誤模擬資料（SIGHAN 與 Wiki）上的表現比較，包含 Match 數、準確率（Acc）與其變化量（ΔAcc）。

爲改善上述問題，我們將這些含語音辨識誤差的資料加入原訓練集，重新訓練出 Model-Mix + NoisyCmd 模型。從表中結果可見，該模型在 Wiki 測試集的準確率從 0.02 顯著提升至 0.44，雖然在 SIGHAN 測試集略降至 0.62，

但整體而言對模型的實用性與穩定性有正向幫助。這說明適當擴增模擬語音輸入錯誤的訓練資料，有助於提升系統在實際語音應用場景中的效能。

綜合上述觀察，若希望系統能更有效地處理語音輸入所帶來的誤差，未來應進一步研究並調整錯誤處理策略，包括擴增訓練資料、設計錯誤修復機制，或結合語音辨識的置信分數等方法，以提升模型在實際語音應用場景中的整體效能。

4 真實使用情境

4.1 建置 API

從上一章的實驗結果，我們已驗證系統架構的可行性。因此我們將本系統設計爲後端 API，允許任何具備語音辨識功能的應用串接使用。同時，我們記錄每次請求的 log，以蒐集更多樣的指令表達方式，協助日後擴充訓練資料。

目前提供兩個開放的 API，其中之一爲 Input Classifier，其流程如圖 6。當使用者透過語音辨識進行輸入時，互動端 (i.e. PC / 手機) 會辨識出語音輸入的結果，接著便將其送至 Input Classifier API，我們會預測出該段內容的類型，並回傳給互動端。如果該段文字內容爲純文字，即可直接顯示在輸入欄，讓使用者判斷該辨識結果是否正確，若該段敘述爲指令，即須呼叫下一個 Error Correction API。

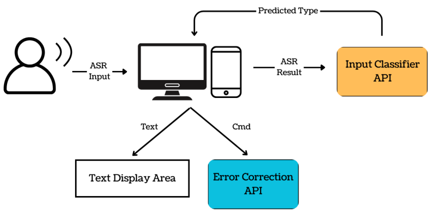


圖 6: Input Classifier API 流程

圖 7 爲 Error Correction API 的整體流程，互動端會將欲修改的句子以及對應的修正指令提供給該 API，我們會先使用 Command Classifier 分類接收到的 command 爲替換、新增和刪除中哪個類別，再透過 Command Labeler 標示出 text 和 command 中，修改位置 (B-Mofify) 以及填入字 (B-Filling) 兩個部分。最後透過 Text Editor 進行修改，直接回傳修改後的內容 (corrected text)。

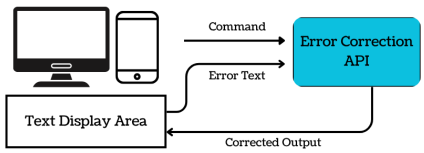


圖 7: Error Correction API 流程

4.2 實際應用

成功建立 API 後，接下來就是實際投入使用。這裡我們以 EduACT [8] 爲例，這是一個提供教師創作對話機器人的平台，協助學生課後學習以及進度追蹤。我們將系統串接至該平台的語音辨識，當使用者使用語音辨識產生內容後，便會將辨識結果傳送給 Input Classifier API，判斷使用者的意圖。舉例來說，某篇文章的名稱爲「數位勞動蟻民」，但使用語音辨識的話，結果會變成「數位勞動移民」。從表 10 的 log 結果也能看出，Input Classifier 的判斷並沒有錯誤。

項目	內容
輸入文字	數位勞動移民
預測結果	0
回傳	`POST /input_classifier/` → 200

表 10: Case Study: 系統 log (Text)

這就是中文語音辨識實際容易遇到的錯誤情況之一，爲了應對這個狀況，使用者即可透過語音辨識下達指令，例如: 「把移民的移改成螞蟻的蟻」。可以看到表 11 的 log 紀錄，我們的系統也能正確的辨認出該文字內容屬於指令，因此回傳結果爲 1，接著便會透過 Error Correction API，判斷指令類型以及進行錯誤標記。

項目	內容
輸入文字	把移動的移改成螞蟻的蟻。
預測結果	1
回傳	`POST /input_classifier/` → 200

表 11: Case Study: 系統 log (指令分類器)

接收到請求之後，我們會先利用 Command Classifier 判斷該指令的類型，如表 12 所示，Command Classiifer 判斷這是一個替換指令，因此 command type 爲 0。接著便是串接欲修改的句子以及指令之後，由 Command Labeler 標記出錯誤位置 (B-Modify) 以及修改字

[8] https://educact.csie.ncu.edu.tw/

詞 (B-Filling) 的位置。最後結合這些預測結果，讓 Text Editor 直接進行修正。

項目	內容
輸入文字	數位勞動移民
使用者指令	把移動的移改成螞蟻的蟻。
預測指令類型	0（替換）
BIO 標註預測	['O', ..., 'B-Modify', ..., 'B-Filling', 'O']
修正後文字	數位勞動蟻民
回傳	POST /error_correction/ → 200

表 12: Case Study: 系統 log (指令標註器)

結果如表12中「修正後文字」欄位所示，修正爲「數位勞動蟻民」，成功完成我們的語音修正任務。透過蒐集每個模型預測的 log 資料，也能擴充我們現有資料集的多元性，對於未來的訓練有很大的幫助。

4.3 修正速度比較

爲了更全面地評估本系統在實際應用中的效益，本節將聚焦於「修正速度」的比較。我們關注的不僅是修正的正確性，也包括修正所需的時間成本。因此，我們設計了一項實驗，針對人爲使用鍵盤修正與透過語音指令修正兩種方式，進行平均耗時的比較。

我們從 SIGHAN-15 以及 zh-tw-wikipedia 的測試集中修正成功的案例中各挑選了 10 筆資料，共 20 筆來進行測試。每筆資料的內容包括:(1) 錯誤句子 (2) 正確句子 (3) 修正指令，如表 13 所示，錯誤句子中最後的「化」字，應該改成「花」。

項目	內容
錯誤句子	對不起，我不能參加你開的慶祝會，因爲我有事情。但是我一定送給你一把很大的化。
修正指令	請把化學的化改成花朵的花
修正結果	對不起，我不能參加你開的慶祝會，因爲我有事情。但是我一定送給你一把很大的花。

表 13: 修正速度測試資料範例

我們分別比較了兩種修正方式。第一種爲人工修正：使用者會看到每筆資料中的錯誤句子與正確修正結果，並透過滑鼠與鍵盤移動至錯誤位置進行編輯。爲聚焦於修正所需時間，我們事先標示出錯誤句與正確句之間的差異，使實驗參與者能直接定位錯誤並執行修正，無須額外花費時間尋找錯誤位置。

第二種方法爲本研究提出的語音指令修正系統。我們提供錯誤句子與對應的修正指令，並記錄從實驗參與者以語音講述指令，到系統完成修正並返回結果的總耗時。儘管部分情況下修正可能無法一次完成，本實驗僅考量成功修正所需的時間，並排除其他可能干擾準確性的因素。

實驗結果如表 14 所示，使用鍵盤輸入進行修正的平均耗時爲約 3 秒，而本研究所提出之語音修正系統的平均耗時爲約 5 秒。儘管語音修正在時間上略高，但差距相對有限，顯示本系統在修正效率上具備實用性與競爭力。

此外，在行動裝置等輸入條件受限的情境中，鍵盤輸入所需的操作時間可能進一步增加。對於不熟悉鍵盤輸入的使用者而言，透過語音指令進行修正可提供更直覺且便利的互動方式，進一步提升系統的可用性。

方法	平均修正時間 (秒)
人工修正	3
語音修正（Model-Mix）	5

表 14: 不同修正方式的平均時間比較

5 結論和未來展望

總結而言，本研究提出一套創新的語音辨識錯誤修正方法，透過口語指令的方式，提升修正彈性與人機互動的自然度。整體系統架構簡潔，僅需輸入分類器、指令分類器與指令標註器三個模組，皆可基於 BERT 系列模型完成訓練，有效降低開發與部署成本。實驗結果亦驗證了各模組的優異表現，展現本系統的實用性與擴展潛力。

此外，我們也利用現有的文字資料集，透過 TTS 以及 ASR 產生語音辨識的錯誤資料集，並建立不同敘述風格的指令類型。我們也以 API 的方式，提供我們的系統給其他具有語音辨識的服務使用，並將蒐集到的眞實資料整理成日後進一步提高模型效能的訓練資料。除此之外，大語言模型的加入也可以讓系統辨識出一些操作類型的指令，擴展使用語音辨識進行互動的靈活性。

儘管如此，系統仍存在若干限制。例如訓練資料多聚焦於單一詞語的單次修改，當句子中錯誤較多時，使用者需連續輸入多筆指令，可能影響修正效率；另一方面，若語音辨識本身輸出錯誤指令內容，亦可能影響最終的修正結果。針對此類問題，未來可嘗試導入其他自動錯誤修正技術，以提升整體系統的穩定性。

針對多語言環境的應用，若能針對各語言的常見辨識錯誤與語句特性，建立專屬的訓練資

料集，則本系統架構亦有望擴展至其他語種，實現跨語言的語音錯誤修正能力。

展望未來，隨著語音互動技術的持續普及，使用者逐漸由鍵盤輸入轉向語音控制。如何自然且有效地修正語音辨識錯誤，將成爲關鍵挑戰之一。本研究所提出之系統，提供語音操作使用者一種新穎且直覺的互動方式，未來在教育、醫療、智慧助理等多元場域皆具備廣泛應用潛力。

參考資料

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Ling Dong, Wenjun Wang, Zhengtao Yu, Yuxin Huang, Junjun Guo, and Guojiang Zhou. 2024. Pronunciation guided copy and correction model for asr error correction. *International Journal of Machine Learning and Cybernetics*, 15(10):4787–4799.

Jin Jiang, Xiaojun Wan, Wei Peng, Rongjun Li, Jingyuan Yang, and Yanquan Zhou. 2024a. Cross modal training for asr error correction with contrastive learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12246–12250.

Lai Jiang, Hongqiu Wu, Hai Zhao, and Min Zhang. 2024b. Chinese spelling corrector is just a language learner. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6933–6943.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, pages 707–710. Soviet Union.

Yinghui Li, Shang Qin, Haojing Huang, Yangning Li, Libo Qin, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, and Philip S Yu. 2024. Rethinking the roles of large language models in chinese grammatical error correction. *arXiv preprint arXiv:2402.11420*.

Yuang Li, Xiaosong Qiao, Xiaofeng Zhao, Huan Zhao, Wei Tang, Min Zhang, and Hao Yang. 2025. Large language model should understand pinyin for chinese asr error correction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Linfeng Liu, Hongqiu Wu, and Hai Zhao. 2024. Chinese spelling correction as rephrasing language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18662–18670.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5054–5065, Hong Kong, China. Association for Computational Linguistics.

Victor Junqiu Wei, Weicheng Wang, Di Jiang, Yuanfeng Song, and Lu Wang. 2024. Asr-ec benchmark: Evaluating large language models on chinese asr error correction. *arXiv preprint arXiv:2412.03075*.

Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. 2025. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. *arXiv preprint arXiv:2501.14350*.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Shuai Zhang, Jiangyan Yi, Zhengkun Tian, Ye Bai, Jianhua Tao, Xuefei Liu, and Zhengqi Wen. 2021. End-to-end spelling correction conditioned on acoustic feature for code-switching speech recognition. In *Interspeech*, pages 266–270.

6 附錄

6.1 指令生成提示詞

```
你是一個文字校對助手，專門協助修改語音辨
    識錯誤的句子。我會提供「正確答案」和
    「題目」，你需要：
1. 比較「正確答案」和「題目」，根據規則
    決定是否需要修改。
   - 如果兩句意義相同，告訴我「不需修
    改」。
   - 如果有需要修改的地方，列出所有修改
    指令，每次只修改一個字。
2. 指令需考慮到用語音辨識輸入時的可辨識
    性，且包含：
   - 指令類型（替換）。
   - 修改位置（「題目」中的第幾個字開
    始）。
   - 被修改的字描述（使用常用詞語，如
    「洗碗的洗」或「希望的希」）。
   - 替換字描述（使用常用詞語，如「辛苦
    的辛」）。
   - 修改後的題目和答案。
3. 修改規則：
   - 忽略標點符號的錯誤。
   - 如果字不同但不影響意思，不算錯字。
   - 比較數字時，統一轉爲阿拉伯數字。
4. 如果有多處需要修改，請列出所有修改指
    令，並保持順序，一次一個字。

#### 範例：
正確答案：我眞的希望我可以去看你。
題目：我眞的洗碗我可以去康你。
指令：
1. 請把『洗碗的洗』改成『希望的希』。
   指令類型：替換
   修改位置：第4個字
   填入字：希
   題目：我眞的{洗}碗我可以去康你。
   答案：我眞的希碗我可以去康你。
2. 請把『洗碗的碗』改成『希望的望』。
   指令類型：替換
   修改位置：第5個字
   填入字：望
   題目：我眞的希{碗}我可以去康你。
   答案：我眞的希望我可以去康你。
3. 請把『健康的康』改成『看見的看』。
   指令類型：替換
   修改位置：第11個字
   填入字：看
   題目：我眞的希望我可以去{康}你。
   答案：我眞的希望我可以去看你。
```

基於寫作風格的圖神經網路假新聞偵測模型
(A Fake News Detection Model Utilizing Graph Neural Networks to Capture Writing Styles)

Yen-Tsang Wu **Lawrence Y. H. Low** **Jenq-Haur Wang**

Department of Computer Science and Information Engineering
National Taipei University of Technology
Web Information Retrieval Lab
t107599005@ntut.edu.tw t113999402@ntut.org.tw jhwang@ntut.edu.tw

摘要

本文提出 CWSMN（Capture Writing Style Multi-Graph Network），一個以圖神經網路為基礎的早期假新聞偵測方法，透過捕捉寫作風格克服傳統語意內容與傳播特徵方法在標註稀缺與跨域泛化不足下的限制。CWSMN 結合文體分析、語意嵌入與多圖融合：以 Bi-GRU 進行上下文初始化，採用 GAT 進行注意力導向的圖聚合，並以 LDA 建構主題圖，最終以輕量級前饋分類器輸出。於多個資料集之實驗顯示，CWSMN 對比 BERT、ALBERT 與 GraphSAINT 等強基準皆有優勢；在未知來源的 Source-CV 場景尤為顯著，證明其於低資源與跨領域環境之穩健泛化能力，並實現不依賴傳播的早期偵測，實驗結果證實本方法在樣本稀缺與未知來源條件下，仍能達成有效的早期偵測。

Abstract

We present CWSMN, a graph neural network for early fake-news detection that foregrounds writing style to address the fragility of purely semantic- or propagation-based approaches under label scarcity and domain shift. CWSMN fuses stylistic cues with semantics through a multi-graph design: Bi-GRU initializes contextual token representations; GAT performs attention-driven aggregation over style- and relation-aware graphs; LDA induces a topic graph, and a lightweight feed-forward head produces predictions. Across multiple datasets, CWSMN consistently surpasses strong baselines (BERT, ALBERT, GraphSAINT), with the largest margins under source-level cross-validation (Source-CV) on unseen sources. These results demonstrate robust generalization in low-resource, cross-domain scenarios and support propagation-agnostic early decisions, underscoring practical value for timely mitigation across platforms and domains.

關鍵字：假新聞偵測、早期偵測、圖神經網路、寫作風格、多圖融合

Keywords: Fake news detection, Graph neural networks, Multi-graph fusion, Early detection, Writing style

1 Introduction

隨著社群媒體與線上新聞平台的蓬勃發展，資訊的傳播速度與規模遠超過以往。然而，這樣的資訊環境同時為假新聞的快速擴散提供了溫床。假新聞所引發的負面影響不僅包括公共輿論的操縱與社會信任的侵蝕，更可能導致經濟損失與公共安全危機(Yang & Pan, 2021)。在近年的重大事件中，如疫情與選舉期間，假新聞所造成的廣泛誤導與社會混亂，進一步凸顯了有效偵測與早期抑制其傳播的迫切需求(Shahid et al., 2022)。傳統的假新聞偵測方法主要可分為兩類：其一為**語意內容導向方法**(Przybyla, 2020)，透過自然語言處理技術抽取文本特徵以進行分類；其二為**傳播結構導向方法**(Cheng et al., 2024)，利用社群媒體互動與訊息擴散模式來辨識真假資訊。雖然這些方法已取得一定成效，但仍面臨幾項挑戰：（1）跨領域的新聞資料存在語言風格與主題差異，使得預訓練語言模型難以有效泛化(Abdali & Krishnamachari, 2022)；（2）標註資料的取得成本高昂且數量有限，導致模型訓練受到限制(Deng & Wang, 2022; Gao et al., 2023)；（3）依賴傳播路徑的模型需等待樣

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 286–295
November 20-22, 2025 ©2025 Association for Computational Linguistics

本累積，無法滿足假新聞「早期偵測」的需求(Deng & Wang, 2022; Shahid et al., 2022)。值得注意的是，雖然假新聞的內容會因領域或事件而有所不同，其**寫作風格**卻往往具有一致性，例如情緒化的詞彙使用、誇張的語氣與特定的句法結構。因此，若能設計能夠有效捕捉寫作風格的模型，即可避免對傳播資訊的依賴，並在訊息發佈初期即進行偵測，達到更即時且跨領域的假新聞辨識效果。基於此，本研究提出一種**多圖寫作風格捕捉網路（Capture Writing Style Multi-Graph Network, CWSMN）**，其核心概念為透過圖神經網路對文本進行多層次風格特徵萃取與融合，進而訓練分類器進行判斷。本研究的主要貢獻可歸納如下：1. **提出寫作風格驅動的假新聞偵測框架**：不同於依賴內容或傳播的傳統方法，本研究專注於捕捉風格特徵，以提升早期偵測能力。2. **設計多圖嵌入與融合策略**：結合 GRU、GAT 與 LDA 生成多種異質子圖，並提出文件層級與詞層級的圖融合方法，有效整合多樣化的風格訊號。3. **驗證跨領域泛化能力**：實驗結果顯示，CWSMN 在未見過來源中顯著優於現有方法，展現了強大的跨領域遷移能力。綜上所述，本研究以寫作風格為核心設計假新聞偵測模型，不僅提升了模型在資料稀缺與跨領域場景下的效能，更具備實務應用價值，能有效支援即時的假新聞擴散。

2 Literature Review

本章回顧與本研究直接相關之理論與方法，包括 Graph Neural Networks 於假新聞偵測的應用、多模態與混合式偵測路徑、寫作風格與文字特徵的假新聞偵測。近年來，GNNs 憑藉其建模結構關係與多模態的能力，逐步成為假新聞偵測的重要技術路線；其中圖注意力網路（Graph Attention Network, GAT）(Veličković et al., 2017)、GraphSAINT(Zeng et al., 2019) 與主題模型（LDA）(Blei et al., 2003)等模型，提供了從詞彙、主題、文件多層次的特徵提取與融合訊的方法，亦為以寫作風格為核心的早期偵測框架奠定基礎。

2.1 圖神經網路於假新聞偵測之基礎與應用

圖神經網路透過訊息傳遞聚合鄰域特徵，能以結構化的方式同時處理文本、主題、來源等異質關係。其中， GAT 以可學習的注意力權重為不同鄰接節點賦予差異化重要度，提升跨語言與結構之表徵能力；結合專為社群和內容而設計的資料集與建模框架，能在訓練資料與測試來源不一致時維持穩健性。其中，多圖架構與注意力機制雖能提升特徵擷取能力，但亦帶來訓練與計算成本。而圖採樣與叢集切分等方法(Chen et al., 2018; Chiang et al., 2019; Dhawan et al., 2024; Goldani et al., 2021; Golovin et al., 2025)使模型在維持辨識能力的同時兼顧記憶體占用與延遲，更能貼近早期偵測與實務部署(Alghamdi et al., 2024; Chang et al., 2024; Phan et al., 2023; Y. Zhang et al., 2024)。

2.2 多模態與混合式假新聞偵測方法

多模態模型能整合文本、視覺與傳播路徑的特徵；混合式方法則結合語意表徵與結構表徵。深度語言模型如 BERT 或 ALBERT 提供強語意表示，但在跨主題或跨來源時，常面臨域偏移問題；與此同時，圖取樣與子圖學習可在大規模圖上有效訓練並保留關鍵關係訊號；兩者互補可提升跨域泛化(Alghamdi et al., 2022; Galli et al., 2022; Mahmoudi et al., 2024; H. Zhang et al., 2024).

2.3 基於寫作風格與文字特徵的假新聞偵測

相較於主題內容，寫作風格在不同領域間更具穩定性，如情緒化詞彙、誇飾語氣、句法節奏與詞彙多樣性。近年工作結合 CoreNLP 的 POS/NER 與 General Inquirer（GI）之心理語意類別以形成細粒度風格訊號，並在圖上建模不同的方式上，以增進來源未知情境下的辨識力（Horne & Adalı, 2018; Potthast et al., 2018; Lima et al., 2020; Stone et al., 1966; Gehrmann et al., 2019）。而寫作風格導向研究顯示，僅依賴語意內容或傳播路徑的偵測器對新興來源與主題漂移較為敏感；反之，將文體作為重要特徵並結合主題資訊與圖結構，可以在跨來源情境取得更穩健表現；相關資料集與基準研究提供來源多樣性與可比性脈絡（Przybyła, 2020; Shu et al., 2018; Galli et al., 2022）。主題建模（如 LDA）能提供文件層級潛在語意結構，適合用於建立 Topic node 與主題感知劃分策略；在模型評估上，除 Doc-CV 外，Topic-CV 與 Source-CV 更能反映真實部署環境，其中 Source-CV 模擬未知來源的泛化能力，常見語意導向模型在此退化，而風格導向加圖融合有較好的效能(Masciari et al., 2020; Nadeem et al., 2023; Nan et al., 2024; Sharma et al., 2023; Tsai, 2023; Wu et al., 2024; Yang et al., 2024)

3 Methodology

為了解決傳統假新聞偵測方法過度依賴人工特徵與語意預訓練模型的限制，我們提出一個名為「基於捕捉寫作風格多圖神經網路的假新聞偵測（CWSMN）」的新穎框架。該模型是基於圖神經網路對於寫作風格的捕捉，旨在將常用的寫作特徵進行提取與融合，應用於假新聞偵測。我們的方法著重於跨領域寫作風格的一致性運用，使得在無需等待使用者傳播模式累積資料的情況下，即可進行早期偵測。CWSMN 包含四個主要階段：風格分析、嵌入、跨模態圖融合與分類。詳細架構如圖 1 所示。

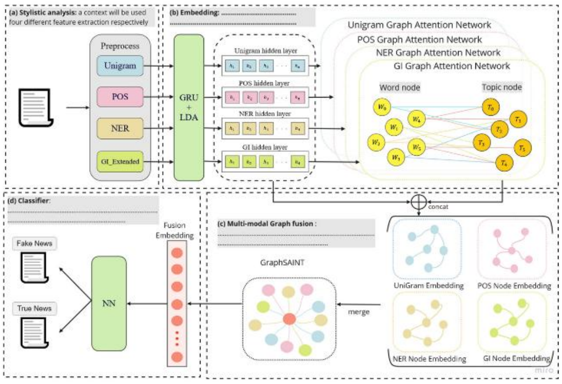


圖 1: CWSMN 架構圖

3.1 Stylistic analysis

在寫作風格分析的整體架構方面，我們採用了一系列文體特徵，進而使用圖注意力機制捕捉寫作風格。在本研究中，我們採用 NER(Named Entity Recognition), POS tagging, and Unigram 以及 General Inquirer dictionaries 作為文件特徵。為了避免分類器過於擬合於特定來源或主題的特徵，在 *N*-gram 的特徵擷取中，我們僅使用詞性單詞（unigrams），不使用雙詞（bigrams）和三詞（trigrams）。我們首先使用 Stanford CoreNLP(Manning et al., 2014) 對輸入文件進行預處理，處理步驟包括 tokenization、NER 以及 POS tagging。除此之外，我們也使用 General Inquirer dictionaries(Stone et al., 1962) 進行正規化並且將所有 word 歸入 182 類別。General Inquirer dictionaries 是一種辭典工具，經常被用於極端黨派新聞識別(Potthast et al., 2017)，字典一共包含 8640 個詞彙。有別於先前的研究 (Przybyla, 2020)，我們並未使用詞向量例如 Word2Vec 來擴充 GI 詞典，因為這類擴充可能會引入語意上的雜訊。

3.2 Embedding and Multi-Graph Construction

我們的詞嵌入過程整合了 Bi-GRU、 GAT 和 LDA，以提取上下文表示和多關係圖結構。嵌入的過程主要分為兩個步驟：嵌入初始化和多圖構建。**Step 1. Embedding Initialization:** 由於 Stylistic analysis 例如 *unigram* 以及 NER 之後的 token，我們將其視為一個新單詞，透過使用雙向 Bi-GRU 根據前後文的關係計算出每個 word vector 作為新 node 的 initialization vector. 這樣的好處是可以賦予每個 node 更多前後文的關係，捕捉更多寫作風格上的特徵。**Step 2: Multi-Graph Construction:** 為了捕捉多樣化的寫作風格關係，我們建構四種以 GAT 為主的圖結構：

- Topic-based Graphs：先以 LDA 模型為每一個 token 計算主題分布，據此為各主題建立主題節點（topic nodes），並將每個 token 與其對應的主題節點連結。藉由這些主題節點，相同主題的 tokens 得以間接關聯。對於同一份文件，我們依不同建構觀點可得到三張圖，分別為 Unigram 圖、NER 圖與 POS 圖。
- GI-based Graphs：依據 GI 字典所定義的語義類別，建立 182 個類別節點；凡屬於同一 GI 類別的詞彙，皆透過對應的類別節點彼此連結，從而強化風格與語用傾向的特徵。

為了捕捉節點之間不同程度的關聯性，我們使用 GAT 來捕捉寫作風格中潛藏的語意與結構。GAT 在進行圖中節點資訊聚合時，引入了注意力機制，使模型能夠根據鄰近節點的重要性，動態學習節點之間的重要性權重。透過注意力機制，模型能自動關注於對風格判別較具代表性的詞彙與其關聯，有效強化跨主題、跨領域的風格一致性建模能力。因此，我們提出的方法透過 Bi-GRU 捕捉到前後文序列的向量，採用 LDA model 與 GI 產生不同結構的 graph，有效強化跨主題、跨領域的風格一致性建模能力。

3.3 Multi-graph Fusion

為了整合來自不同圖結構的資訊，我們採用了兩種融合策略。

3.3.1 Token-level Fusion

為建構全域層級（global graph）的語意關係圖，我們將每個 token 視為節點，並為每份 document 建立一個 document node；該文件內之所有 token-nodes 與其 document node 連邊。考量圖結構規模龐大，訓練成本較高，我們引入 GraphSAINT 演算法進行子圖抽樣，以提升模型的訓練效率與可擴展性。GraphSAINT 透過基於節點、邊或隨機游走的子圖抽樣技術，僅從原始大圖中選取部分子圖進行訓練，既保留了全圖的結構資訊，又有效降低了運算成本。如下圖所示。

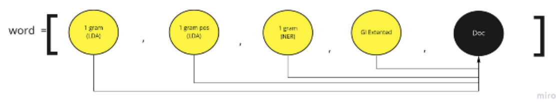


圖 2: Token-level 融合示意圖。

GraphSAINT 是一種針對大規模圖神經網路訓練所提出的高效子圖抽樣方法。GraphSAINT 透過基於節點、邊或隨機游走的子圖抽樣技術，僅從原始大圖中選取部分子圖進行訓練，既保留了全圖的結構資訊，又有效降低了運算成本。與其他 mini-batch 訓練方法相比，GraphSAINT 在保留圖結構統計特性方面表現更佳，並能在不犧牲精度的前提下，大幅提升訓練效率。因此，由於實驗的 tokens 的數量十分龐大，因此我們使用 GraphSAINT 作為我們全域層級的融合模型。經由 GraphSAINT 後，我們可以獲得每一個 document 的 embedding。

3.3.2 Document-level Fusion

另外一種融合方式為 Document level fusion。我們分成 2 個步驟。

- Embedding aggregation：對四張子圖的詞嵌入採平均池化以得其子圖表示（subgraph representation）。
- GAT-based fusion: 使用 GAT 模型進行 fusion。 我們創立一個 document node，然後 document node 與四個子圖 node 連接，進行運算以融合這些資訊。GAT 通過注意力機制動態計算文件節點與各子圖表示節點之間的權重（注意力分數），以決定每個子圖表示對最終文件表示的貢獻。最終更新文件節點的表示，融合來自四個子圖的特徵，形成局部層級(Local Graph)的文件表示。如下所示:

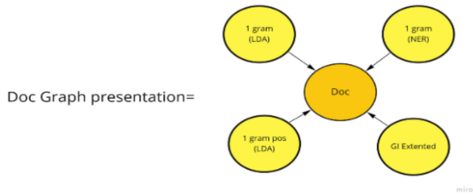


圖 3: Document-level 融合示意圖。

由於 document-level fusion 只使用 5 個 node。因此，我們使用 GAT 模型，通過注意力機制動態地為圖中的節點分配不同的權重，從而捕捉節點間的關係和重要性。最後，我們獲得每一個 document node 的 embedding 作為融合後 document 的特徵向量。

3.4 Classification

通過圖融合（graph fusion，使用 GAT 融合四個子圖的表示，生成反映局部層級的文件向量）後，將得到的文件向量輸入到一個簡單的前饋神經網路（feed-forward neural network, FNN）分類器，進行二元分類，判斷新聞是真（real）還是假（fake）。Pseudo-code 如下:

Algorithm 1 Training CWSMN

```
Require: corpus D, graphs {G^(g)}, parameters θ
Ensure: trained parameters θ
D ← PreprocessWithCoreNLP(D)        ▷ tokenize, GI/POS/NER; LDA with K=100
E_word ← InitWordEmbeddingsWithGRU(D);    E_nonword ← InitLearnableNonWordVectors()
{G^(g)} ← BuildGraphs(D, {TW, CO, SP, EN, GI, DW})
for epoch = 1 to E do        ▷ Word-level path on DW with GraphSAINT
    for sampled G_sub ⊂ G^(DW) via GraphSAINT do
        H^(DW) ← GNN_Forward(G_sub)
        z̃_d^(w) ← READOUT(H^(DW); center = d)
    end for
                                  ▷ Graph-level path on {TW, CO, SP, EN, GI}
    for g ∈ {TW, CO, SP, EN, GI} do
        H^(g) ← GAT_2Layers(G^(g))
        z̃_d^(g) ← AttentionAvg({ h_w^(g) | w ∈ W(d) })
    end for
    {z_d^(g)}_g ← SoftmaxGate({z̃_d^(g)}_g)        ▷ gated cross-graph fusion
    z_d ← Proj(Concat(z_d^(w), z_d^(g)))
    p̂ ← Softmax(FFN(z_d))
    L ← L_cls(p̂, y_d) + α L_align + λ L_sparse
    θ ← AdamW_Update(θ, ∇_θ L)
end for
```

4 Experiment

4.1 Dataset description

我們使用一個公開的 dataset 來評估我們提出的 CWSMN 模型。

4.1.1 資料特徵

實驗用的資料集來自(Przybyla, 2020)從網路爬取的一個語料庫，內容包含 52,790 篇的假新聞與 50,429 篇的真新聞，總計 103,219 篇文本文件、超過 1.17 億個詞元（tokens），一共 205 個不可信網站與 18 個可信網站作為資料來源。為提升語料品質，我們排除了重複內容與那些平均每行少於 15 個詞彙的文本，以確保文本具有足夠語意密度。我們觀察資料集特性

可以發現，假新聞來源較常涉及的主題包括穆斯林與移民、健康與營養、總統競選對手比較等；而真新聞則較常涵蓋的主題則為電影與體育等。

4.1.2 資料集劃分方式

為了評估模型在不同情境下的表現，實驗設計了三種場景進行交叉驗證 Document-based CV，Topic-based CV 和 Source-based CV。

- Document-based CV: 將所有 103,219 篇文件隨機分成 5 個 fold，確保每個 fold 包含來自不同來源和主題的文件。模擬測試文件來自已知來源和已知主題的情況，作為基準評估。
- Topic-based CV: 使用潛在狄利克雷分配生成 100-topic LDA 模型，並將每篇文件分配到與其關聯最強的主題。將這些主題隨機分成 5 個 fold，確保每個 fold 包含與特定主題相關的所有文件。目的是為了模擬真實世界裡文件屬於訓練資料中未見過的主題的情況，測試模型對新主題的泛化能力。這種劃分方式確保訓練和測試資料的主題分離，防止模型依賴特定主題的詞彙或內容進行分類。
- Source-based CV: 將所有 223 個來源分成 5 個 fold，確保每個折包含來自特定來源的所有文件。每個 fold 的測試集包含來自訓練資料中未見過的來源的文件。目的是模擬真實世界假新聞來自全新來源（例如新興新聞網站）的情況，測試模型對新來源的泛化能力。

實驗的資料集劃分策略旨在模擬現實世界中假新聞檢測的挑戰，特別是新主題和新來源的情境。通過隨機文件劃分、主題分離和來源分離，實驗確保模型不僅在已知資料上表現良好，還能泛化到未知情境。

4.2 Implementation Details

GAT 模型架構所有參數都遵循 Graph Attention Networks 論文裡的設定，Number of Attention Heads 為 8. Number of Layers 為 2. 激活函數分別為 LeakyReLU；輸出層使用 softmax。GraphSAINT 根據相同的文件路徑構建圖結構，每個 node 的維度設為 100。Activation function 採用 LeakyReLU。使用 Adam 作為優化器。在圖抽樣過程中，我們使用隨機游走抽樣（random walk sampling）方法。分類損失使用 cross-entropy loss. 為了捕捉不同新聞來源間的主題差異，我們利用潛在狄利克雷分配訓練一個包含 100 個主題的模型。此後，將每個詞彙根據其最大主題關聯度，指派至對應的主題。

4.3 Comparison of Methods

我們將本模型與當前最先進的方法進行比較以評估其效能，並同時測試多個模型變體：

- BERT(Devlin et al., 2019): BERT 是一種基於 Transformer 的雙向語言模型，廣泛用於自然語言處理任務。受限於最大序列長度，在本研究中，僅使用文件的前 512 個 token。
- ALBERT(Lan et al., 2019): 是一種相較於 BERT 參數量更少的模型，並且專注於句子間的連貫性。通過參數縮減技術和自監督學習目標，顯著降低模型大小，同時保持高性能。兩者皆屬於預訓練的 Transformer 模型，透過擷取文字語意來進行分類。
- Word2Vec +GraphSAINT: 透過 Word2Vec 計算出 word 的 embedding 後，相加取平均作為 Document 的向量。 然後依據傳播路徑建圖進行分類。
- ALBERT+GraphSAINT: 先透過 ALBERT 計算出所有 token 的 embedding 後相加取平均作為 Document 的向量。然後依據傳播路徑建圖進行分類。
- GI+GraphSAINT: 僅使用 GI 字典裡的字向量相加取平均作為 Document 的向量.然後依據傳播路徑建圖進行分類。 以上三種方法都是根據傳播路徑進行假新聞偵測，屬於圖神經網路。
- Stylometric: Stylometric 分類器是一種基於文體特徵的模型，專注於捕捉文本的寫作風格，避免依賴特定來源或主題特徵。該方法通過提取文體相關特徵並使用線性模型進行分類。
- BiLSTMAvg: 是一種基於雙向長短期記憶網絡的神經網絡模型，通過對文件中所有句子的表示進行平均，生成文件級別的預測。該模型旨在捕捉句子的上下文和文體特徵，用於假新聞檢測的文體分析。

- Bag-of-Words: 是一種簡單的基線分類器，基於詞彙頻率表示文件，用於假新聞檢測。以上三種方法都是假新聞偵測常用的方法，採用傳統文字語意作為特徵-包含統計詞彙頻率與文體分析。

4.4 Evaluation metric

由於真和假新聞的數量大致平衡，因此我們實驗使用準確率（accuracy）作為評估指標。此外，為了確保每個來源和主題的文件都被用於測試，我們採用 5-fold cross-validation。

4.5 Experimental Results

本節呈現了我們在假新聞檢測任務中的實驗結果，比較了多種模型在三種交叉驗證情境下的表現：文件級交叉驗證、主題級交叉驗證和來源級交叉驗證。實驗結果如下表:

Model	Doc-CV	Topic-CV	Source-CV	Average
CWSMN_GAT	0.9870	0.9760	0.9784	0.9805
CWSMN_GraphSAINT	0.9930	0.9910	0.9746	0.9862
ALBERT	0.9815	0.9754	0.7165	0.8911
ALBERT+GraphSAINT	0.9985	0.9961	0.7385	0.9110
Word2Vec+GraphSAINT	0.9993	0.9853	0.9095	0.9647
GI+GraphSAINT	0.9983	0.9995	0.7096	0.9025
Stylometric	0.9274	0.9173	0.8097	0.8848
BiLSTMAvg	0.8994	0.8921	0.8250	0.8722
Bag-of-Words	0.9913	0.9886	0.7078	0.8959
BERT	0.9976	0.9965	0.7960	0.9300

表 1：各模型在三種場景下的準確率。

Document CV scenario: Document cv 模擬已知來源和主題的場景，測試的文件來自已知來源和主題。結果顯示：Word2Vec+GraphSAINT 表現最佳，達到 0.9993 的準確率，這表示複合類型的假新聞偵測模型-結合語意和傳播路徑有最強的效能。而預訓練模型 BERT 和 ALBERT+GraphSAINT、GI+GraphSAINT 達到 0.9993 準確率，優於其餘方法，顯示在已知資料上的優勢。而我們提出的模型效能雖然輸給最好的模型，分別為 0.0115 跟 0.0063，但是也能高達 0.987 與 0.993，遠勝過於傳統的方法。傳統的方法中 Stylometric 和 BiLSTMAvg 表現相對較弱，可能是因為文體特徵提取的簡單性限制了其在已知資料上的表現。**Topic CV scenario:** Topic-CV 模擬新事件場景，測試文件來自未見過的主題（基於 LDA 分配的 100 個主題）。結果顯示：GI+GraphSAINT 達到最高準確率 0.9995，顯示其對新主題的強適應性，歸功於 GI 詞典的情感特徵和 GraphSAINT 的結構化建模。我們提出的模型分別達到 0.9910 與 0.9760 略輸最好的模型 0.0085 跟 0.0235。 值得一提的是 Bag-of-Words (0.9886) 僅下降 0.27%（相較於 Doc-CV 的 0.9913），顯示詞彙的頻率對主題變化的有效性。而傳統的方法表現略低，表明文體特徵對主題變化的適應能力較為不足。

Model	Doc-CV	Topic-CV	Source-CV	Average
BERT	0	0	0	0
ALBERT fine-tune	-0.0161	-0.0211	-0.0795	-0.0389
CWSMN_GAT	-0.0106	-0.0205	0.1824	0.0504
CWSMN_GraphSAINT	-0.0046	-0.0055	0.1786	0.0562

Source CV scenario: Source-CV 模擬新興假新聞網站場景，測試文件來自未見過的來源，是最嚴苛的泛化測試。結果顯示：我們所提出的兩個模型表現最佳，分別為 0.9784 跟 0.9746，這顯示結合寫作風格特徵和圖神經網路的方法具有卓越泛化能力。在這個 scenario 裡，我們的方法領先了基於傳播路徑與語意的多模態模型 Word2Vec+GraphSAINT (0.9095)分別為 0.0689 與 0.0651.而領先 pretrained Transformer models such as BERT (0.7960)、ALBERT (0.7165)分別達到 0.2619 與 0.1824.這證明了比起大成本的預訓練模型，我們的方法能夠以較小的計算量達到更好的效果。此外，我們觀察到，BiLSTMAvg (0.8250) 和 Stylometric (0.8097) 在文體特徵模型中表現最佳，這也更加證明，使用以寫作風格作為特徵的方法，更符合真實世界中，對新興假新聞的有效性。

平均準確率: 平均準確率綜合評估模型在三種情境下的整體性能。我們提出的兩種方法都達到最佳的效果(0.9862 與 0.9805)，顯示其在不同 scenario 下的均衡表現，擁有優異的性能。而多模態模型 Word2Vec+GraphSAINT (0.9647) 表現出色，僅輸給我們的模型 0.0751 pretrained Transformer models 分別落後我們 0.056 與 0.095，受 Source-CV 準確率較低的影響。整體而言，實驗結果顯示我們所提出 CWSMN 模型在未曾學習過的資料上情境下展現出優異的假新聞偵測能力。，在已有資料可以學習的情境下，也有不俗的效能，相比於預訓練模型或是圖神經網路，我們的平均準確率最高，最高可達 0.9862，進一步證明了本模型的強大效能。

5 Evaluation and Analysis

5.1 與預訓練的深度語意模型的比較

表 2：不同預訓練的深度語意模型的比較結果

BERT 與 ALBERT fine-tune 都是使用雙向 Transformer 編碼器的預訓練模型，我們 fine-tune 之後進行預測，模型性能以 BERT 為基準，差值表示相對於 BERT 的準確率變化。Doc-CV 場景下，ALBERT fine-tune 相較於 BERT 下降 0.01606，可能是由於參數縮減（12M-235M 相較於 BERT 的 110M-340M）影響了模型的效能。CWSMN_GAT 則下降 0.0106，表明基於特徵組合的模型在已知資料上與 BERT 接近，但效能略遜。CWSMN_GraphSAINT 下降最少，顯示 GraphSAINT 的採樣技術增強了特徵表達能力，接近 BERT 的性能。Topic-CV 場景下，ALBERT fine-tune 相較於 BERT 下降 0.0211，表明其對新主題的適應性略弱，可能是 SOP 目標未完全補償 NSP 的影響。CWSMN_GAT 則下降 0.0205，與 ALBERT 接近，CWSMN_GraphSAINT 下降最少，顯示 GraphSAINT 的結構化建模有效提升了對新主題的泛化能力。Source-CV 模擬新興假新聞網站場景，測試資料來自未見過的來源，這在早期偵測中，是最重要的一點。ALBERT fine-tune 相較於 BERT 下降 0.0795，而我們提出的 CWSMN_GAT 和 CWSMN_GraphSAINT 相較於 BERT 分別提升 0.1824 跟 0.1786，表明基於寫作風格在捕捉文體特徵方面優於預訓練模型。BERT 和 ALBERT fine-tune 作為基準模型，在已知樣本中效能優秀，但在 Source-CV 由於未曾學習相關的深度語意，因此限制了泛化能力，準確率只有 0.796。而我們提出的方法，主要依靠寫作風格作為特徵，不受未知主題與語意的限制，對新來源的顯著提升，顯示結合單詞、命名實體、詞性標記和 GI 詞典的情感特徵能有效捕捉文體模式，超越預訓練的深度語意模型，這證明了對於現今不斷新增類型與主題的動態假新聞偵測上的重要性。

5.2 Comparison of Graph Models with Different Edge Construction Methods

Model	Doc-CV	Topic-CV	Source-CV	Average
Word2Vec_GraphSAINT	0	0	0	0
ALBERT+GraphSAINT	-0.0008	0.0108	-0.171	-0.0537
GI_GraphSAINT	-0.001	0.0142	-0.1999	-0.0622
CWSMN_GAT	-0.0123	-0.0093	0.0689	0.0158
CWSMN_GraphSAINT	-0.0063	0.0057	0.0651	0.0215

表 3：不同建邊策略之比較

本節比較了採用不同建邊方法的圖模型（Graph Models）在假新聞檢測任務中的性能，特別聚焦於不同特徵的比較。表 3 列出了五種圖神經網路模型在三種場景下的準確率變化。在 document-CV 與 Topic-CV 下，我們的方法的準確率僅些微落後於最強基準；然而在最具挑戰性的 Source-CV 情境中，本方法取得最佳表現。此結果顯示：當測試分布出現來源轉換時，「以寫作風格為特徵並透過多圖融合」的設計能帶來更好的跨來源泛化能力。同樣基於圖神經網路，但 GraphSAINT 的建邊機制核心在於：若兩個文件來自同一網站，則在圖上形成較強連結。此設計等同於將來源特徵視為主要特徵。然在現實中，虛假資訊網站往往快速出現又迅速消失，且有大量短命或一次性域名；因此，即時偵測時常無法依賴「過去的資料」來推斷真偽。結果是：在 Source-CV 情境下，過度依賴來源同質性的建邊策略容易失效，並導致對新來源的遷移能力不足。我們亦嘗試將來源關係納入本模型（將來源感知的建邊與多圖風格圖結合，類似於與 GraphSAINT 的混合）。實驗顯示，整體效能反而下降。可能原因包括：**overfitting：**風格圖已提供可區辨的文體訊號；再疊加來源同質性，容易學到「來源＝標籤」的虛假關聯，降低對新來源的魯棒性。**訊號冗餘與過度同質（over-smoothing）：**來源邊會把同站點文件過度拉近，使得節點表示在圖上趨於同質，削弱內容/風格細節。**分布偏移（distribution shift）：**訓練時的來源分布與測試差異過大，來源邊成為不穩定的遷移支點。

5.3 與 Baseline 的比較

Model	Doc-CV	Topic-CV	Source-CV
Stylometric	0	0	0
BiLSTMAvg	-0.028	-0.0252	0.0153
Bag-of-Words	0.0639	0.0713	-0.1019
BERT	0.0702	0.0792	-0.0137
CWSMN_GAT	0.0596	0.0587	0.1687
CWSMN_GraphSAINT	0.0656	0.0737	0.1649

表 4：Baseline 的比較結果

首先，相較於傳統文本方法（如 BoW、Stylometric、統計式分類器等），本研究方法在三種場景設定上皆取得全面領先，顯示以寫作風格為特徵的多圖融合能夠在已知與未知分

布下同時維持高準確率與穩健性。其次，相較於同為神經網路的基準模型，本研究方法仍呈現顯著優勢。這證明：單純提升網路性能不足以克服跨來源的分布轉移；反之，結合多視角建邊與跨子圖注意力融合，更能提取對假新聞具有域不變性的風格線索。第三，就詞袋模型而言，本研究方法在 document-CV 與 Topic-CV 皆呈小幅領先，而在 source-CV 上的優勢更為明顯，領先幅度達 0.2706。此一差距反映：當測試來源為未曾出現時，依賴來源或主題近鄰的建邊策略容易失效；相對地，以文體／語用為核心的多圖表示能更有效對抗來源漂移。第四，與目前最具代表性的深度語意模型 BERT 相比，雖然本研究方法在 Document-CV 與 Topic-CV 略遜一籌，但在 Source-CV 卻展現顯著領先，優勢高達 0.1824。此結果凸顯：僅依賴語意預訓練的表徵在面對新來源時較易退化；相較之下，風格驅動＋多圖融合提供了更具泛化性的決策訊號。

5.4 Ablation Study

我們以下列四類特徵分別建圖並以 GAT 進行聚合：Unigram， NER， POS， GI，並且測試在上述四特徵融合下，再疊加 GraphSAINT 的來源建邊之影響。其結果如下表。

Model	Doc-CV	Topic-CV	Source-CV
GI	0.9728	0.9671	0.7179
POS+GI	0.9733	0.966	0.7449
NER+GI	0.9726	0.9645	0.7317
Unigram+NER	0.9864	0.98	0.9456
Unigram+GI+POS	0.9862	0.986	0.9624
Unigram+NER+POS	0.9872	0.984	0.958
Unigram+NER+POS+GI (Average)	0.987	0.976	0.9784
Unigram+NER+POS+GI+GraphSAINT	0.993	0.991	0.9746

表 5： Ablation Study 結果。

僅使用 GI 建圖能在 Doc-CV 與 Topic-CV 達到合理表現，但在 Source-CV 顯著落後，說明僅靠心理語意類別不足以支撐面對新來源的泛化。而 POS+GI 與 NER+GI 在三種 CV 均優於單獨 GI，代表結合句法或實體脈絡能補足 GI 的粒度；但在 Source-CV 的提升仍有限，顯示僅雙特徵尚不足以克服來源漂移。再加入 Unigram 的三種特徵組合後在三種 CV 均有明顯增益，尤其 Source-CV 提升最為突出，顯示 Unigram 對未知來源的區辨尤為關鍵。將 4 種特徵以平均方式融合，在 Doc-CV 與 Topic-CV 維持高準確率，同時在 Source-CV 取得全表最佳或近最佳；此設定提供了「已知分布準確」與「未知來源泛化」間的最優特徵。但是，值得注意的，4 種特徵融合基礎上加入來源同站建邊後，Doc-CV 與 Topic-CV 小幅上升，但 Source-CV 略為降低，這證明來源邊容易讓模型學到來源＝標籤的誤差，對未見來源產生過擬合。由結果顯示，在三種 CV 下，四訊號融合取得最佳整體表現，特別是在 Source-CV 顯示最強泛化；而在此基礎上再加入來源建邊，雖能微幅提升 Doc/Topic-CV，卻不利於 Source-CV，印證來源特徵對未知來源的過擬合風險。

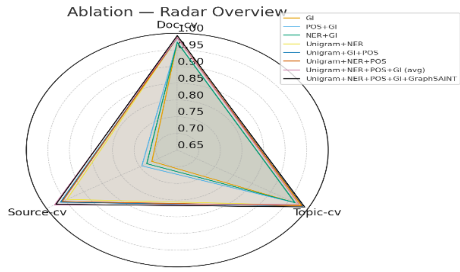


圖 4: Ablation 效果總覽

6 Conclusion

本研究提出 CWSMN 模型，以寫作風格作為偵測核心，透過寫作風格等多視角建邊，採用兩種融合策略：Token-level（以 GraphSAINT 子圖抽樣訓練）與 Document-level（以 GAT 進行跨子圖注意力融合），在三種場景進行測試與消融分析，得到以下結論：1.跨來源泛化的優勢：在 Source-CV 的嚴苛情境，CWSMN 顯著優於各組對照，對目前最強深度語意模型亦具明確領先，驗證「風格驅動結合多圖融合」能有效對抗來源漂移並支撐早期偵測。2. 已知分布下的穩健性：在 Doc 以及 Topic-CV 中，CWSMN 與最佳比較模型效能相當或僅小幅落後；顯示在不犧牲已知分布準確的情況下，仍能換取對未知來源的泛化能力。3.早期偵測的可能性: 由於不依賴傳播軌跡，因此可以達到早期假新聞偵測的目的。

Acknowledgments

作者感謝國家科學及技術委員會（NSTC）之經費支持（計畫編號：NSTC 114-2221-E-027-068、NSTC 114-2634-F-027-001-MBK），使本研究得以順利進行。亦感謝匿名審查者提供的中肯建議，對於論文品質之提升助益良多；同時感謝參與資料蒐集、系統建置與實驗驗證

之人員。本文內容僅代表作者個人觀點與責任，與補助機關立場無涉。

References

Abdali, S., & Krishnamachari, B. (2022). Multi-modal misinformation detection: Approaches, challenges and opportunities. *arXiv preprint arXiv:2203.13883*.

Alghamdi, J., Lin, Y., & Luo, S. (2022). Modeling fake news detection using bert-cnn-bilstm architecture. 2022 IEEE 5th international conference on multimedia information processing and retrieval (MIPR),

Alghamdi, J., Luo, S., & Lin, Y. (2024). A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, *83*(17), 51009-51067.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Chang, W., Liu, K., Yu, P. S., & Yu, J. (2024). Enhancing Fairness in Unsupervised Graph Anomaly Detection through Disentanglement. *arXiv preprint arXiv:2406.00987*.

Chen, J., Ma, T., & Xiao, C. (2018). Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv preprint arXiv:1801.10247*.

Cheng, L.-C., Wu, Y. T., Chao, C.-T., & Wang, J.-H. (2024). Detecting fake reviewers from the social context with a graph neural network method. *Decision Support Systems*, *179*, 114150.

Chiang, W.-L., Liu, X., Si, S., Li, Y., Bengio, S., & Hsieh, C.-J. (2019). Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining,

Deng, Y., & Wang, S.-W. (2022). Detecting Fake News on Social Media by CSIBERT. Proceedings of the 2022 6th International Conference on Deep Learning Technologies,

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),

Dhawan, M., Sharma, S., Kadam, A., Sharma, R., & Kumaraguru, P. (2024). Game-on: Graph attention network based multimodal fusion for fake news detection. *Social Network Analysis and Mining*, *14*(1), 114.

Galli, A., Masciari, E., Moscato, V., & Sperlí, G. (2022). A comprehensive Benchmark for fake news detection. *Journal of Intelligent Information Systems*, *59*(1), 237-261.

Gao, W., Ni, M., Deng, H., Zhu, X., Zeng, P., & Hu, X. (2023). Few-shot fake news detection via prompt-based tuning. *Journal of Intelligent & Fuzzy Systems*, *44*(6), 9933-9942.

Goldani, M. H., Momtazi, S., & Safabakhsh, R. (2021). Detecting fake news with capsule neural networks. *Applied Soft Computing*, *101*, 106991.

Golovin, A., Zhukova, N., Delhibabu, R., & Subbotin, A. (2025). Improving Recommender Systems for Fake News Detection in Social Networks with Knowledge Graphs and Graph Attention Networks. *Mathematics*, *13*(6), 1011.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mahmoudi, G., Behkamkia, B., & Eetemadi, S. (2024). Zero-Shot Stance Detection using Contextual Data Generation with LLMs. *arXiv preprint arXiv:2405.11637*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations,

Masciari, E., Moscato, V., Picariello, A., & Sperli, G. (2020). A deep learning approach to fake news detection. International Symposium on Methodologies for Intelligent Systems,

Nadeem, M. I., Ahmed, K., Zheng, Z., Li, D., Assam, M., Ghadi, Y. Y., Alghamedy, F. H., & Eldin, E. T. (2023). SSM: Stylometric and semantic similarity oriented multimodal fake news detection.

Journal of King Saud University-Computer and Information Sciences, *35*(5), 101559.

Nan, Q., Sheng, Q., Cao, J., Hu, B., Wang, D., & Li, J. (2024). Let Silence Speak: Enhancing Fake News Detection with Generated Comments from Large Language Models. *arXiv preprint arXiv:2405.16631*.

Phan, H. T., Nguyen, N. T., & Hwang, D. (2023). Fake news detection: A survey of graph neural network methods. *Applied Soft Computing*, *139*, 110235.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Przybyla, P. (2020). Capturing the style of fake news. Proceedings of the AAAI conference on artificial intelligence,

Shahid, W., Jamshidi, B., Hakak, S., Isah, H., Khan, W. Z., Khan, M. K., & Choo, K.-K. R. (2022). Detecting and mitigating the dissemination of fake news: Challenges and future research opportunities. *IEEE Transactions on Computational Social Systems*.

Sharma, A., Sharma, M., & Dwivedi, R. K. (2023). Exploratory data analysis and deception detection in news articles on social media using machine learning classifiers. *Ain Shams Engineering Journal*, *14*(10), 102166.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*(4), 484.

Tsai, C.-M. (2023). Stylometric fake news detection based on natural language processing using named entity recognition: In-domain and cross-domain analysis. *Electronics*, *12*(17), 3676.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wu, J., Guo, J., & Hooi, B. (2024). Fake news in sheep's clothing: Robust fake news detection against LLM-empowered style attacks. Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining,

Yang, H.-C., Hung, Y.-L., & Wang, L.-C. (2024). Stylometry-based Fake News Classification Using Text Mining Techniques. Proceedings of the 2024 11th Multidisciplinary International Social Networks Conference,

Yang, J., & Pan, Y. (2021). COVID-19 Rumor Detection on Social Networks Based on Content Information and User Response. *Frontiers in Physics*, *9*, 763081.

Zeng, H., Zhou, H., Srivastava, A., Kannan, R., & Prasanna, V. (2019). Graphsaint: Graph sampling based inductive learning method. *arXiv preprint arXiv:1907.04931*.

Zhang, H., Liu, X., Yang, Q., Yang, Y., Qi, F., Qian, S., & Xu, C. (2024). T3RD: Test-Time Training for Rumor Detection on Social Media. Proceedings of the ACM on Web Conference 2024,

Zhang, Y., Ma, X., Wu, J., Yang, J., & Fan, H. (2024). Heterogeneous Subgraph Transformer for Fake News Detection. Proceedings of the ACM on Web Conference 2024,

Revisiting Pre-trained Language Models for Conversation Disentanglement

Tung-Thien Lam
Dept. of Computer Sci. and Eng.
Yuan Ze University
Taoyuan, Taiwan
s1136058@mail.yzu.edu.tw

Cheng-Zen Yang
Dept. of Computer Sci. and Eng.
Yuan Ze University
Taoyuan, Taiwan
czyang@saturn.yzu.edu.tw

Abstract

Multi-party conversation is a popular form in online group chatting. However, the interweaving of utterance threads complicates the understanding of the dialogues for participants. Many conversation disentanglement models have been proposed using transformer-based pre-trained language models (PrLMs). However, advanced transformer-based PrLMs have not been extensively studied. This paper investigates the effectiveness of six advanced PrLMs: BERT, XLNet, ELECTRA, RoBERTa, DeBERTa, and ModernBERT. The experimental results show that DeBERTa has outstanding performance than other PrLMs for the conversation disentanglement task.

Keywords: Multi-party Conversation, Conversation Disentanglement, Pre-trained Language Models, Performance Evaluation

1 Introduction

Online group chatting provides important channels to multiple participants to communicate, discuss opinions, and share information. Due to its popularity, a huge amount of conversation is generated daily. Since multiple participants are present in a chatting room simultaneously, there are many different utterance threads of various topics concurrently happening in the room and they are usually intertwined without specific structural information. Although these dialogues contain abundant valuable information, the interweaving of utterance threads complicates the understanding of the dialogues for participants (Shen et al., 2006; Elsner and Charniak, 2010; Uthus and Aha, 2013). Figure 1 shows a simplified example in which only two interwoven utterance threads are illustrated. In the dialogue, the utterance threads lack coherence not only for they are intertwined but also for many irrelevant utterances appear between these threads.

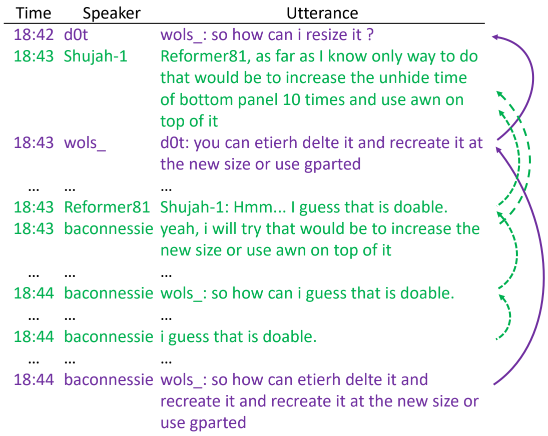


Figure 1: An example of two utterance threads extracted from the Ubuntu IRC data (Kummerfeld et al., 2019). They are expressed in purple and green.

Recently, many conversation disentanglement models (Zhu et al., 2020, 2021; Li et al., 2022; Ma et al., 2022) have proposed by employing pre-trained language models (PrLMs) such as BERT (Devlin et al., 2018) and ALBERT (Lan et al., 2020) to improve the disentanglement performance. However, previous research has not extensively explored the effectiveness of advanced transformer-based models such as ELECTRA (Clark et al., 2020) and DeBERTa (He et al., 2021).

In this paper, six advanced transformer models are investigated for their effectiveness in conversation disentanglement: BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2020), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and ModernBERT (Warner et al., 2025). To evaluate the performance of each PrLM, we construct the disentanglement model based on the MF (Manual Features) model (Zhu et al., 2021) because the kernel of MF is relatively concise by using a 2-layer FFN (Feed-Forward Network) model. Different PrLM models can be simply employed in MF and evaluated. The experiments are con-

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 296–302
November 20-22, 2025

ducted using the Ubuntu IRC dataset (Kummerfeld et al., 2019) because this dataset has been widely used to evaluate the disentanglement performance of different approaches.

The constructed MF-based model is a two-step disentanglement model. The first step is to perform link prediction to find the reply-to relation between a target utterance and a group of history utterances. Based on the link prediction results, the second step is to perform clustering to determine the utterance threads. The experimental results show that DeBERTa outperforms other PrLMs in terms of both link prediction metrics and clustering metrics.

The rest of the paper is organized as follows. Section 2 reviews previous studies employing PrLMs on conversation disentanglement. Section 3 describes the task definition and the dataset. Section 4 describes the studied pre-trained transformer-based models. Section 5 presents the experiments and discusses the experimental results. Finally, Section 6 concludes the paper.

2 Related Work

Prior studies have proposed various models for the multi-party conversation disentanglement problem (Uthus and Aha, 2013)(Uthus and Aha 2013). As pre-trained language models (PrLMs) have been widely used in natural language processing (NLP) tasks, many recent disentanglement models employ PrLMs to improve the disentanglement performance. In 2020, Zhu et al. proposed a masked hierarchical transformer model (Zhu et al., 2020) using BERT to generate feature vectors and make pairwise decisions. In 2021, Zhu et al. studied three transformer-based PrLMs with the MF model (Zhu et al., 2021): BERT (Devlin et al., 2018), ALBERT (Lan et al., 2020), and Poly-Encoder (Humeau et al., 2020). Their experimental results show that BERT combined with MF outperforms other models.

In 2022, Ma et al. proposed a BERT-based model StructBERT considering structural information of dialogues (Ma et al., 2022). In StructBERT, BERT is used to capture the contextual information of utterances. Li et al. proposed a hierarchical pre-trained model DialBERT (Li et al., 2022) using BERT to capture the matching relationship between two utterances. Jiang et al. proposed an intent-based mutual learning model MuiDial (Jiang et al., 2022) using BERT to generate utterance embeddings. In 2023, Bhukar et al. proposed an end-to-end deep reinforcement learning model (Bhukar et al., 2023) using StructBERT to get high-quality link prediction results.

In 2024, Gao et al. proposed an end-to-end implicit addressee model IAM (Gao et al., 2024) using BERT to generate utterance embeddings. Li et al. proposed a model using discourse-aware encoding and hierarchical ranking loss (DiHRL) (Li et al., 2024). As StructBERT, DiHRL uses BERT to perform contextual information encoding for utterances.

To the best of our survey, only the work of Zhu, Lau, and Qi Zhu et al. (2021) have investigated three PrLMs. This paper investigates more advanced transformer-based PrLMs that have been proposed recently.

3 Task Definition and Datasets

Since this work investigates the effectiveness of various PrLMs based on MF disentanglement model (Zhu et al., 2021), this paper frames the task as a problem to find reply-to relations (link prediction) and discover utterance threads (clustering). Given an utterance u_i in a dialogue D and a list of candidate prior utterances $\{u_j\}$ in the same dialogue, the disentanglement model firstly predicts the parent utterance of u_i from $\{u_j\}$. After all reply-to relations in a segment of D have been predicted, the model performs clustering to aggregate utterance threads.

To evaluate the effectiveness of PrLMs, the Ubuntu IRC dataset (Kummerfeld et al., 2019) is used because it has been widely used for performance evaluation in many studies. This dataset consists of three parts: 67,463 utterances for training, 2,500 utterances for validation, and 5000 utterances for testing.

4 Pre-Trained Models

To evaluate these PrLMs, we construct an MF-based model in which a pairwise model is used to predict the reply-to link relations as shown in Figure 2, where k_h defines the number of utterances including the target utterance u_i for reply-to relevance calculations, w_r^t is the t-th word embedding in the r-th utterance, and mf_{ij} represents the manually defined features including the utterance characteristics and the mutual relationships like the number of intervening messages, the word overlap ratio, and the condition of words in common. In this work, k_h is set to 50. Thereafter, our MF-based

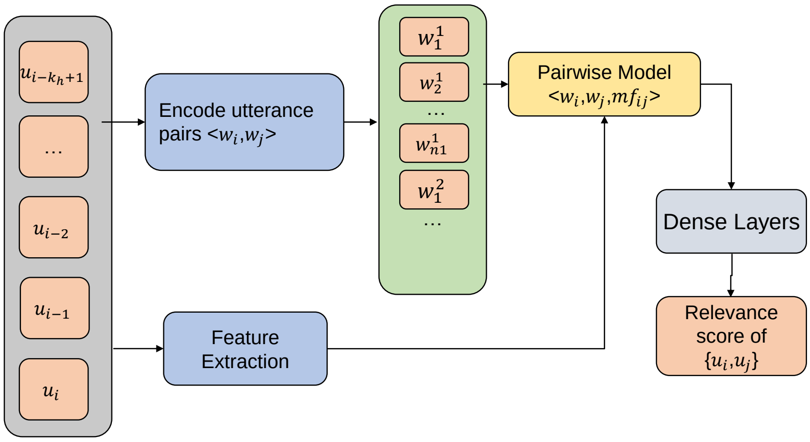


Figure 2: The pairwise model to perform link prediction in the MF-based model.

model uses the Union-Find algorithm as (Kummerfeld et al., 2019) instead of the bipartite graph algorithm used in (Zhu et al., 2021) to perform clustering because of the wide employment of Union-Find in many related studies.

In this paper, six PrLMs are investigated. They are listed as follows:

- **BERT**: This model uses Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) to read a given sentence bidirectionally. This enables BERT to capture both left and right context simultaneously, improving the understanding of semantics. It has been widely used in many disentanglement models.

- **XLNet**: This model employs an autoregressive pre-training method to learn context bidirectionally. As shown in (Yang et al., 2019), XLNet outperforms BERT on many NLP tasks.

- **ELECTRA**: This model employs a replaced token detection method instead of an MLM approach. It trains a discriminator to distinguish real tokens from the replaced ones, and uses a MLM-based generator to predict corrupted tokens. This allows ELECTRA to use all tokens of the given input for learning, gaining more computational efficiency and parameter-effectiveness than BERT.

- **RoBERTa**: This model enhances the performance by modifying several BERT design features, including removing the NSP loss and training on a larger corpus with dynamic masking.

- **DeBERTa**: This model employs two mechanisms, Disentangled Attention and Enhanced Mask Decoder, to enhance task performance. With the Disentangled Attention approach, DeBERTa represents the content and relative position information of a token into two distinct vectors. With the Enhanced Mask Decoder approach, DeBERTa considers the absolute position information of tokens in the decoding layer to capture more complementary information. Compared with RoBERTa-Large, DeBERTa can achieve better performance with less training data.

- **ModernBERT**: This model integrates modern refinements, including rotary positional embeddings, root mean square (RMS) normalization, and multi-query attention into the BERT core. Compared with BERT, ModernBERT is optimized for longer context lengths.

5 Experiments

We have conducted experiments to evaluate the disentanglement performance of the MF-based model using different transformer-based PrLMs. The following implementations are used for the studied PrLMs: bert-base-uncased for BERT, xlnet-base-cased for XLNet, electra-base for ELECTRA, roberta-base for RoBERTa, deberta-v3-base for DeBERTa, and ModernBERT-base for ModernBERT.

We use Precision, Recall, and F1 to measure the link prediction performance. They are defined as follows:

$$\text{Precession} = \frac{\text{Correctly predicted links}}{\text{All predicted links}}, \qquad (1)$$

$$\text{Recall} = \frac{\text{Correctly predicted links}}{\text{All true links}}, \quad (2)$$

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

For clustering performance, we use 1-VI (Variation of Information), ARI (Adjusted Rand Index), MCP (Matched-Cluster Precision), MCR (Matched-Cluster Recall), and MCF (Matched-Cluster F1). Because VI shows the dissimilarity between two clusters, this work uses 1-VI defined as follows:

$$1 - VI = 1 - \frac{H(Y|X) + H(X|Y)}{\log(n)}, \quad (4)$$

where X and Y represent two utterance threads, $H()$ is the entropy function, and n is the number of the utterances. ARI shows the similarity of two clusters according to the links. It is defined as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{l_i}{2} \sum_j \binom{l_j}{2}\right]}{\binom{n}{2}}}{\frac{\left[\sum_i \binom{l_i}{2} + \sum_j \binom{l_j}{2}\right]}{2} - \frac{\left[\sum_i \binom{l_i}{2} \sum_j \binom{l_j}{2}\right]}{\binom{n}{2}}}, \quad (5)$$

where n_{ij} is the number of links that appear in the predicted cluster i and also the true cluster j, l_i is the number of the links in the cluster i, l_j is the number of the links in the cluster j, and n is the number of the ground truth links. MCP is the Precision of the exactly-matched clusters. MCR is the Recall of the exactly-matched clusters. MCF is the harmonic mean of MCP and MCR. They are defined as follows:

$$\text{MCP} = \frac{\text{Exactly matched clusters}}{\text{All predicted clusters}}, \quad (6)$$

$$\text{MCR} = \frac{\text{Exactly matched clusters}}{\text{All true clusters}}, \quad (7)$$

$$\text{MCF} = \frac{2 \times MCP \times MCR}{(MCP + MCR)}. \quad (8)$$

All models are executed 10 times with random initializations on GPUs. The results are averaged.

We use the same settings for all models without any fine-tuning. There are three hidden layers (256, 128, 64). The optimizer is AdamW. The learning rate is 5e-5. The number of epochs is 10. The loss function is CrossEntropyLoss. The dropout rate is 0.1. The batch size is 2 with a gradient accumulation of 32. The maximum number of tokens of an utterance is 60.

Table 1 shows their link prediction performance. From Table 1, we can find that DeBERTa outperforms other PrLMs for link prediction and ELECTRA takes second place. BERT continues to deliver consistent performances. However, ModernBERT does not perform well. One possible reason is that the Ubuntu IRC dialogue dataset is a kind of the QA task, and the utterance threads are interwoven. The characteristics of the Ubuntu IRC dialogue dataset hinder the performance of ModernBERT as the findings in (Antoun et al., 2025).

Model	**Precision**	**Recall**	**F1**
BERT	0.7277	0.7014	0.7144
XLNet	0.7209	0.6951	0.7077
ELECTRA	0.7288	0.7024	0.7153
RoBERTa	0.7281	0.7019	0.7147
DeBERTa	**0.7364**	**0.7100**	**0.7230**
ModernBERT	0.7219	0.6960	0.7087

Table 1: Link prediction performance of the MF-based model with different PrLMs.

Table 2 shows the clustering performance of each model. DeBERTa still outperforms other PrLMs. The results of Tables 1 and 2 show that DeBERTa achieves the best performance among the studied PrLMs.

Model	**1-VI**	**ARI**	**MCP**	**MCR**	**MCF**
BERT	0.9095	0.6217	0.3323	0.3915	0.3594
XLNet	0.9064	0.6222	0.3340	0.3721	0.3518
ELECTRA	0.9123	0.6543	0.3382	0.3865	0.3606
RoBERTa	0.9132	0.6424	0.3312	0.3986	0.3616
DeBERTa	**0.9175**	**0.6644**	**0.3656**	**0.4175**	**0.3897**
ModernBERT	0.9068	0.6128	0.3273	0.3837	0.3529

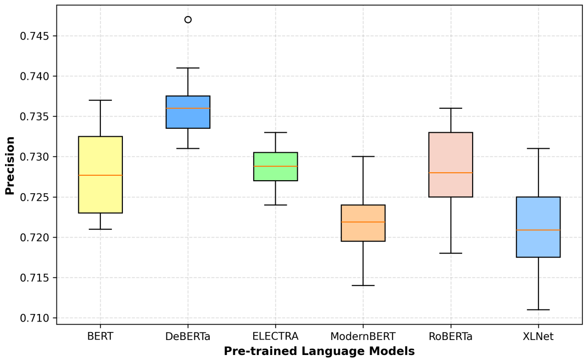
Table 2: Clustering performance of the MF-based model with different PrLMs.

Figure 3 shows the boxplot of the performance of the studied PrLMs in terms of the investigated metrics in the experiments. As shown in Figure 3, DeBERTa also has the best median scores on all performance metrics among the studied PrLMs.

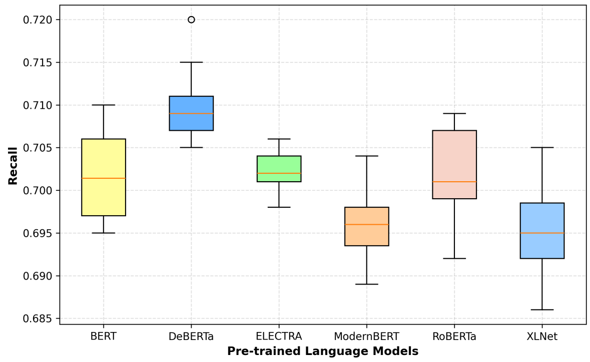
6 Conclusions

Multi-party conversation is a popular form to discuss opinions, share information, and discover solutions for problems. However, the interweaving of utterance threads complicates the understanding of the dialogues for participants.

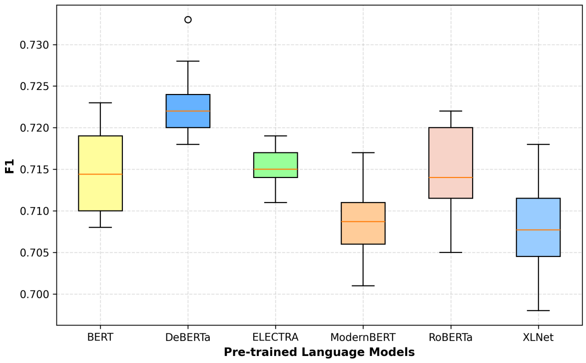
In the past, many conversation disentanglement models have been proposed using transformer-based PrLMs. However, advanced transformer-based PrLMs have not been extensively investigated.



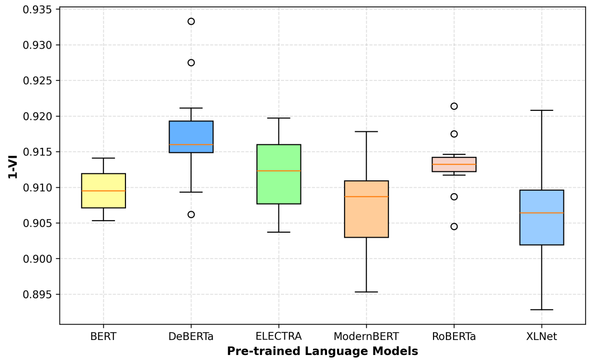
(a) Precision



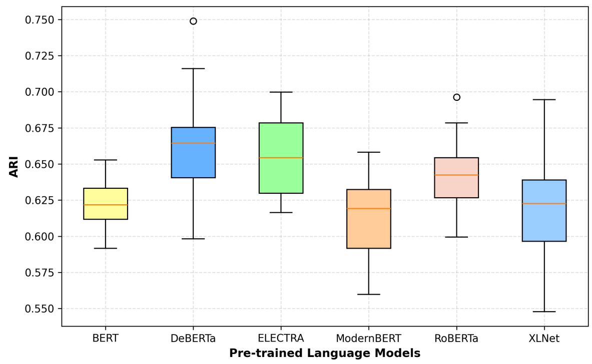
(b) Recall



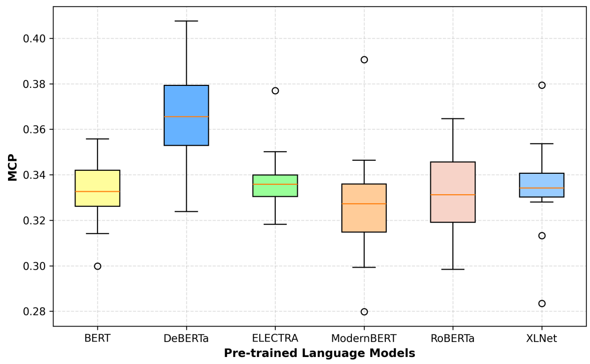
(c) F1



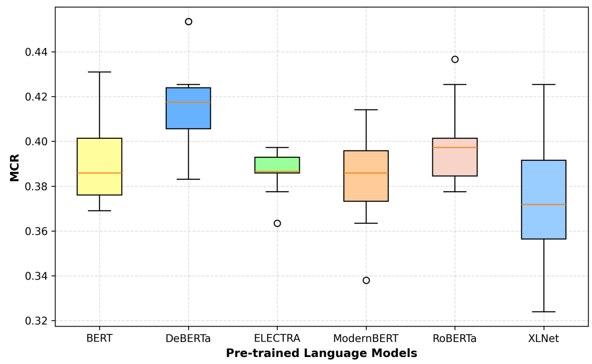
(d) 1-VI



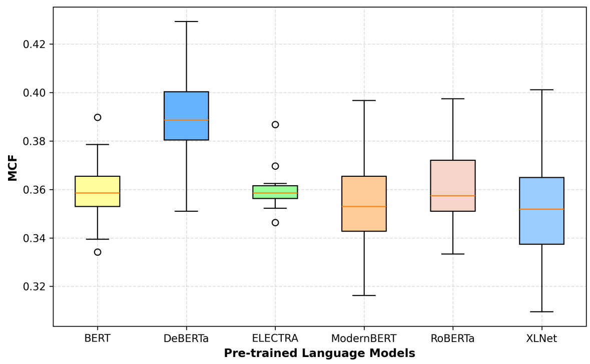
(e) ARI



(f) MCP



(g) MCR



(h) MCF

Figure 3: Boxplot of performance of the MF-based model with different PrLMs.

In this paper, six advanced transformer-based models are investigated for their effectiveness in conversation disentanglement: BERT, XLNet, ELECTRA, RoBERTa, DeBERTa, and ModernBERT. The experimental results show that DeBERTa outperforms other PrLMs for the conversation disentanglement task.

There are still some issues to be investigated further in the future. Firstly, our study does not discuss the best performance of each model because we use the same settings for all models without any finetuning. Therefore, more extensive investigations will be conducted to explore the best performance of these models. Secondly, the number of the studied PrLMs is limited. In the future, other advanced PrLMs will be included in the investigation. Finally, the MF-based model considers the manually defined features that are extracted for the Ubuntu IRC dataset. Other disentanglement models with better generalizability will be considered for more comprehensive analysis on PrLMs.

Acknowledgments

This work is supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC 113-2221-E-155-048-MY2. The authors would also like to express their sincere thanks to anonymous reviewers for their precious comments.

References

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2025. ModernBERT or DeBERTaV3? Examining Architecture and Data Influence on Transformer Encoder Models Performance. *CoRR*, arXiv:2504.08716.

Karan Bhukar, Harshit Kumar, Dinesh Raghu, and Ajay Gupta. 2023. End-to-End Deep Reinforcement Learning for Conversation Disentanglement. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI-23)*, volume 37, pages 12571–12579.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*, pages 4171–4186.

Micha Elsner and Eugene Charniak. 2010. Disentangling Chat. *Computational Linguistics*, 36(3):389–409.

Jingsheng Gao, Zeyu Li, Suncheng Xiang, Zhuowei Wang, Ting Liu, and Yuzhuo Fu. 2024. Toward an End-to-End Implicit Addressee Modeling for Dialogue Disentanglement. *Multimedia Tools and Applications*, 83(28):70883–70906.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021)*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Ziyou Jiang, Lin Shi, Celia Chen, Fangwen Mu, Yumin Zhang, and Qing Wang. 2022. MuiDial: Improving Dialogue Disentanglement with Intent-Based Mutual Learning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 4164–4170.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C. Polymenakos, and Walter Lasecki. 2019. A Large-Scale Corpus for Conversation Disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-seng Chua, and Donghong Ji. 2024. Revisiting Conversation Discourse for Dialogue Disentanglement. *ACM Transactions on Information Systems*, 43(1).

Tianda Li, Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2022. Conversation- and Tree-Structure Losses for Dialogue Disentanglement. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 54–64.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2022. Structural Characterization for Dialogue Disentanglement. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 285–297.

Dou Shen, Qiang Yang, Jian-Tao Sun, and Zheng Chen. 2006. Thread Detection in Dynamic Text Message Streams. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 35–42.

David C. Uthus and David W. Aha. 2013. Multiparticipant Chat Analysis: A Survey. *Artificial Intelligence*, 199–200:106–121.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Henghui Zhu, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Who Did They Respond to? Conversation Structure Modeling Using Masked Hierarchical Transformer. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, 05, pages 9741–9748.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2021. Findings on Conversation Disentanglement. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association (ALTA 2021)*, pages 1–11.

Multilingual Promise Verification in ESG Reports with Large Language Model Performance Evaluation

Wei-Chen Huang
Bachelor of Statistics
National Taipei University
New Taipei City, Taiwan
wesley@gmail.com

Hsin-Ting Lu
Graduate Institute of
Information Management
National Taipei University
New Taipei City, Taiwan
hsintinglubob@gmail.com

Wen-Ze Chen
Graduate Institute of
Information Management
National Taipei University
New Taipei City, Taiwan
50712andy@gmail.com

Min-Yuh Day*
Graduate Institute of
Information Management
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

Abstract

Corporate sustainability reports often contain vague or unverifiable statements, increasing the risk of greenwashing. As global expectations for the credibility of ESG disclosures continue to rise, developing automated systems capable of verifying corporate sustainability commitments has become an important research direction. However, current analytical approaches still face limitations in multilingual ESG promise verification, particularly in non-English language contexts.

This study investigates the performance of a large language model (GPT-5) in cross-lingual ESG promise verification tasks by evaluating corporate reports in Chinese, Japanese, and English, with the goal of establishing a multilingual evaluation benchmark. Four core subtasks are examined, including promise identification, evidence status assessment, evidence quality evaluation, and verification timeline prediction. Multiple prompting strategies—from zero-shot to few-shot learning, including Chain-of-Thought reasoning—are systematically compared to analyze the effectiveness of different design choices.

Results show that few-shot prompting generally yields more stable verification performance, while evidence quality evaluation remains the most challenging task across languages. Theoretically, this study proposes a cross-lingual prompting framework that clarifies how task complexity and annotation imbalance influence LLM reasoning performance in ESG verification. Practically, the findings provide actionable implications for regulators, investors, and corporate decision-makers by supporting the deployment of AI-based monitoring systems to enhance disclosure credibility, strengthen governance resilience, and enable more informed sustainable finance decisions.

Keywords: PromiseEval, Multilingual Dataset, Promise Verification, Greenwashing, Large Language Model

1 Introduction

This study aims to develop an advanced framework leveraging Large Language Model (LLM) to automatically verify whether corporate ESG reports contain explicit promises and to evaluate their credibility. Environmental, Social, and Governance (ESG) reporting has become a cornerstone of corporate accountability, with stakeholders increasingly relying on sustainability disclosures to inform investment decisions and assess corporate responsibility. However, the proliferation of ESG reporting has been accompanied by a concerning rise in greenwashing practices, where corporations overstate their environmental and social commitments while obscuring less favorable activities (Delmas & Burbano, 2011; Lyon & Montgomery, 2015). This phenomenon not only misleads stakeholders but also weakens the credibility of sustainability reporting.

Recent research emphasizes that greenwashing is both common and difficult to measure, as textual claims often lack clear evidence or measurable outcomes (Testa et al., 2018; Wang et al., 2025). To address this issue, computational approaches have been developed to automatically detect sustainability-related commitments and assess their validity. The PromiseEval shared task (Chen et al., 2025) introduced the first multilingual

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 303–313
November 20-22, 2025 ©2025 Association for Computational Linguistics

benchmark for corporate promise verification, defining four subtasks:

1. **Promise Identification (PI):** determine whether a segment expresses promising contents.
2. **Supporting Evidence Assessment:** assess whether promises contain concrete evidence.
3. **Clarity of the Promise–Evidence Pair (CPEP):** evaluate the clarity and relevance of evidence in relation to the promise.
4. **Timing for Verification (TV):** indicate when a promise should be revisited for verification (e.g., within_two_years, two_to_five_years, more_than_5_years, others).

Building upon this foundation, the ML-Promise dataset (Seki et al., 2024) expanded multilingual coverage to five languages and incorporated retrieval-augmented generation techniques, demonstrating the feasibility of cross-lingual promise verification.

Despite progress in ESG analysis, significant analytical gaps persist. A systematic review by Lublóy et al. (2025) underscores the fragmented nature of current greenwashing quantification methods, a problem reflected in several key areas, like cultural and linguistic disparities, lack of integrated pipelines, insufficient verification baseline. Our study directly addresses these gaps by pioneering a multilingual framework that leverages Large Language Models (LLMs).

This study addresses these persistent analytical gaps, particularly concerning Chinese and Japanese reports, through three key contributions: (1) examining the feasibility of promise verification across multiple languages, (2) establishing baseline methods using state-of-the-art large language models for comparative analysis, and (3) providing methodological foundations for automated multilingual ESG verification systems.

2 Literature Review

2.1 ESG Reporting and Greenwashing

Greenwashing has been widely studied in sustainability communication research. Delmas and Burbano (2011) provide a conceptual framework for understanding the drivers of greenwashing, while Lyon and Montgomery (2015) emphasize its prevalence and regulatory implications. Empirical studies confirm that sustainability reports often contain misleading or unverifiable claims (Testa et al., 2018), reinforcing the need for computational tools. More recently, Wang, Gao, Wang et al. (2025) developed a greenwashing index using deep learning, providing quantitative evidence of discrepancies between corporate claims and substantiating evidence.

2.2 Computational Approaches to Detect Greenwashing

Recent advances in text mining and natural language processing (NLP) have been applied to analyze corporate sustainability disclosures and detect potentially misleading claims. For example, Wang Wang et al., (2025) proposed automated greenwashing indices derived from textual features of corporate reports, demonstrating how linguistic signals can indicate discrepancies between promises and actual practices. Beyond domain-specific applications, shared tasks such as SemEval-2022 Task 8 on Multilingual News Article Similarity (Chen et al., 2022) illustrates how NLP benchmarks can evaluate semantic consistency across texts in multiple languages. These computational approaches highlight the potential of AI-driven methods for large-scale monitoring of sustainability communication and for identifying unverifiable or vague ESG-related claims.

2.3 Corporate Promise Verification Tasks

The PromiseEval shared task, introduced at SemEval-2025 (Chen et al., 2025), formally established promise verification as a natural language processing (NLP) challenge. It defined tasks that align closely with the detection of vague or unverifiable claims in sustainability disclosures, focusing not only on promises but also on supporting evidence, clarity, and timeline. Its design highlights the complexity of promise verification and its close relationship to greenwashing detection

2.4 Multilingual NLP Datasets and Benchmarks

Multilingual benchmarks such as ML-Promise (Seki et al., 2024) have extended promise verification to multiple languages, addressing the gap in non-English corporate reporting. Other multilingual resources in NLP, such as XNLI (Conneau et al., 2018), show the value of multilingual evaluation, but ML-Promise is the first domain-specific dataset focused on corporate promises.

2.5 Annotation Quality and Inter-Annotator Agreement

Annotation quality is critical for promise verification tasks. Artstein and Poesio (2008) emphasize the role of inter-annotator agreement (IAA) metrics such as Cohen's Kappa and Krippendorff's Alpha to ensure annotation reliability. Both PromiseEval (Chen et al., 2025)and ML-Promise (Seki et al., 2024) adopted these metrics, reporting substantial agreement levels ($\kappa > 0.6$), which supports the validity of the datasets and subsequent analyses.

3 Experimental Setup

3.1 System Architecture

This study investigates multilingual promise evaluation using large language models as the foundational architecture. Promise evaluation encompasses the identification and verification of commitments within textual content, representing a critical component for assessing corporate statements in Environmental, Social, and Governance (ESG) reporting. Our research examines three linguistically diverse languages (Chinese, Japanese, and English), evaluating model performance across four distinct subtasks: Promise Identification (PI), Evidence Status Assessment (ESA), Evidence Quality Evaluation (EQE), and Verification Timeline Prediction (VTP).

To systematically assess model capabilities, we implement five prompting strategies: zero-shot, one-shot, three-shot, and five-shot learning, as well as an additional five-shot variant enhanced with Chain-of-Thought (CoT) prompting. Referencing the study by (Wei et al., 2022), we believe that Chain-of-Thought (CoT) requires sufficient examples to guide reasoning. Therefore, our analysis focuses on the 5-shot setting, as this configuration is not only our best-performing one, but also because the 5 examples provide sufficient context to allow us to isolate the effect of explicit reasoning. Within this specific setup, we systematically evaluate the marginal benefits of CoT. Building on this setup, the comprehensive evaluation framework enables controlled comparison of how demonstration quantity and reasoning instructions influence classification performance across different languages and verification tasks.(Figure 1 illustrates the overall research framework.)

3.2 Dataset

The study uses the PromiseEval dataset (Seki et al., 2024), which provides multilingual samples annotated for ESG-related promise verification. For each of the three languages (Chinese, Japanese, and English), the dataset is divided into 400 training samples and 400 test samples.

- Promise Status: classification of whether a concrete or organizational-level commitment is present.
- Evidence Status: detection of whether verifiable supporting evidence is provided.
- Evidence Quality: evaluation of evidence clarity (Clear, Not Clear, Misleading, N/A).
- Verification Timeline: identification of the expected timeline of promise fulfillment (Already, Within 2 years, Between 2–5 years, More than 5 years, N/A).

During the preliminary stage, samples were uniformly formatted for input to GPT-5. For few-shot conditions, demonstration examples were randomly sampled from the training set to prevent test data leakage and to simulate realistic evaluation scenarios.

In addition, to provide a clearer understanding of dataset composition, we analyzed the label distributions across the three languages. Table 1 and Table 2 present the distributions of the Chinese, Japanese, and English subsets. Although each training and test set contains the same number of samples (400 each), the label proportions across subtasks remain imbalanced, such as differences between positive and negative samples. These distributional characteristics may influence model classification performance.

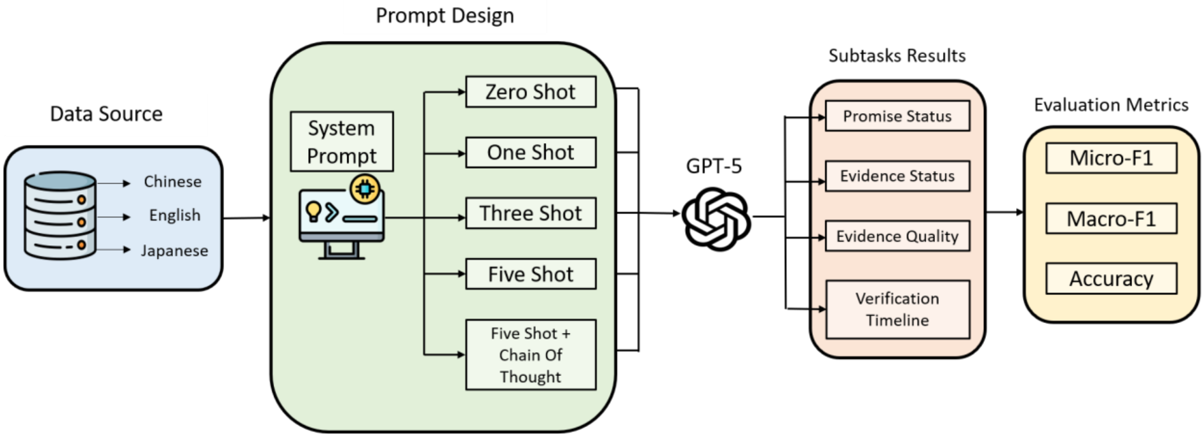


Figure 1: Proposed research workflow for ESG promise verification

Task	Label	Chinese	Japanese	English
Promise Status	Yes	146 (36.50%)	356 (89.00%)	313 (78.25%)
	No	254 (63.50%)	44 (11.00%)	87 (21.75%)
Evidence Status	Yes	78 (19.50%)	279 (69.75%)	221 (55.25%)
	No	322 (80.50)	77 (19.25%)	179 (44.75%)
	N/A	-	44(11.00%)	-
Evidence Quality	Clear	50(12.50%)	161(40.25%)	132(33.00%)
	Not Clear	16 (4.00%)	106 (26.50%)	85 (21.25%)
	Misleading	1 (0.25%)	12 (3.00%0	4 (1.00%)
	N/A	333 (83.25%)	121 (30.25%)	179 (44.75%)
Verification Timeline	Already	-	282 (70.50%)	155 (38.75%)
	Within 2 years	55 (13.75%)	18 (4.50%)	36 (9.00%)
	Between 2 to 5 years	7 (1.75%)	22 (5.50%)	75 (18.75%)
	More than 5 years	29 (7.25%)	34 (8.50%)	47 (11.75%)
	N/A	309 (77.25%)	44 (11.00%)	87 (21.75%)

Table 1: Label distribution of the PromiseEval training datasets (Chinese, Japanese, and English)

Task	Label	Chinese	Japanese	English
Promise Status	Yes	237 (48.47%)	372 (93.00%)	273 (68.25%)
	No	252 (51.53%)	28 (7.00%)	127 (31.75%)
Evidence Status	Yes	148 (30.27%)	232 (58.00%)	206 (51.50%)
	No	341 (69.73%)	140 (35.00%)	194 (48.50%)
	N/A	-	28 (7.00%)	-
Evidence Quality	Clear	73 (14.93%)	142 (35.50%)	134 (33.50%)
	Not Clear	46 (9.41%)	84 (21.00%)	71 (17.75%)
	Misleading	-	6 (1.50%)	1 (0.25%)
	N/A	370 (75.66%)	168 (42.00%)	194 (48.50%)
Verification Timeline	Already	-	295 (73.75%)	143 (35.75%)
	Within 2 years	101 (20.65%)	19 (4.75%)	36 (9.00%)
	Between 2 to 5 years	11 (2.25%)	17 (4.25%)	50 (12.50%)
	More than 5 years	39 (7.98%)	41 (10.25%)	44 (11.00%)
	N/A	338 (69.12%)	28 (7.00%)	127 (31.75%)

Table 2: Label distribution of the PromiseEval test datasets (Chinese, Japanese, and English)

3.3 Model and Strategies

We adopted GPT-5 as the unified Large Language Model (LLM) architecture across all languages and subtasks, focusing on evaluating the impact of prompt-based inference on classification performance. We designed five distinct prompting strategies.

First, regarding Prompting Strategies, we evaluated the following five settings:

- 0-shot: Consisted only of the task definition and system instructions in the prompt.
- 1-shot: The prompt was supplemented with one demonstration example.
- 3-shot: The prompt was supplemented with three demonstration examples.
- 5-shot: The prompt was supplemented with five demonstration examples.
- 5-shot + CoT (Chain-of-Thought): The prompt was supplemented with five demonstrations, along with an additional Chain-of-Thought instruction to encourage step-by-step logical reasoning before the final answer. However, the model was strictly required to output only the final structured label.

Second, concerning the Demonstration Source and Sampling for Few-Shot learning, all demonstration examples were selected from the training subset of the PromiseEval dataset to strictly prevent test data leakage and simulate realistic In-Context Learning evaluation scenarios. Given the significant class imbalance in our dataset (particularly for the Evidence Quality and Verification Timeline tasks), we employed a Stratified Random Sampling mechanism to select demonstrations. This ensured that the class label distribution in each Few-shot prompt (e.g., 'Yes'/'No' for Promise Status) maintained an approximate balance relative to the overall training set. This method aims to provide the LLM with a representative and stable context, thereby mitigating the class bias that pure random sampling might introduce.

This design enables a controlled comparison of how the number of demonstrations (from 0 to 5) and reasoning instructions (CoT) affect classification performance across languages and subtasks.

3.4 Evaluation Metrics

Model predictions on the test sets were compared against gold-standard annotations. Performance was measured using:

- Accuracy: Calculates the proportion of correctly predicted samples over the total number of samples, reflecting overall correctness at the instance level.
- Micro-F1: Aggregates true positives, false positives, and false negatives across all classes, reflecting overall predictive accuracy.
- Macro-F1: Computes F1-scores for each class independently, then averages them, ensuring fair evaluation of minority classes.

Together, these metrics provide a comprehensive assessment of both global accuracy and class-level robustness across multilingual promise evaluation tasks.

4 Experiment Results and Analysis

This section presents the performance evaluation of the GPT-5 model across the PromiseEval subtasks and provides an interpretative analysis within the context of the SemEval-2025 Task 6 shared task results. Our analysis covers performance metrics across three languages (Chinese, Japanese, and English) and five distinct prompting strategies, aiming to establish robust benchmarks for multilingual promise verification.

4.1 Overall Performance Analysis

Our results confirm that few-shot prompting consistently outperformed the zero-shot baseline, aligning with the principles of effective in-context learning. Table 3 shows that the 5-shot configuration is optimal, yielding the highest mean Accuracy (71.12%) and Macro-F1 (51.92%) across all tasks and languages. Conversely, the incorporation of Chain-of-Thought (CoT) reasoning led to a marginal and statistically non-significant decrease in aggregate performance (Accuracy 70.58%; Macro-F1 51.04%). However, a consistent downward trend was observed across multiple subtasks, suggesting that explicit reasoning did not consistently benefit pattern-based classification. A more detailed analysis of this phenomenon is provided in Section 4.4.

Strategy	Accuracy	Macro-F1	Micro-F1
0 shot	69.54%	47.66%	69.54%
1 shot	70.46%	49.95%	70.46%
3 shot	70.81%	51.37%	70.81%
5 shot	**71.12%**	**51.92%**	**71.12%**
5 shot_COT	70.58%	51.04%	70.58%

Table 3: Overall Performance Across All Tasks and Languages (Mean Values).

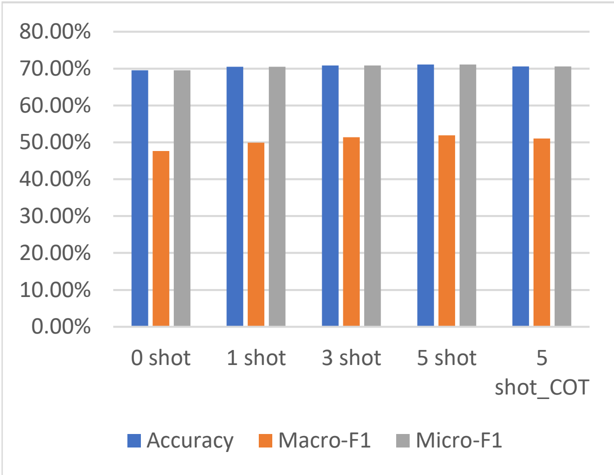


Figure 2: Overall Performance Across All Tasks and Languages (Mean Values).

4.2 Task-Specific Performance

4.2.1 Promise Status Identification

Promise Status Identification emerged as the most tractable subtask among those evaluated. Table 4 shows high performance across all languages, with the zero-shot setting in Japanese reaching 92.25% accuracy.

However, this high accuracy is a misleading artifact of severe class imbalance, where Table 1 shows the "Yes" class constitutes 89.8% of the Japanese dataset. This imbalance allows the model to achieve an inflated score by simply defaulting to the majority prediction. Consequently, the accuracy metric fails to penalize the model for its poor performance on the minority "No" class, a well-documented issue known as the Accuracy Paradox.

In contrast, the Macro F1 score provides a more robust evaluation by mitigating this bias. It achieves this by calculating the F1 score (which balances Precision and Recall) for each class independently before computing an unweighted average, thus giving equal importance to both majority and minority classes. Under this more reliable metric, the 5-shot setting emerges as the top performer for Japanese with a Macro F1 score of 73.12%, outperforming the zero-shot setting's score of 71.64%. This trend, where multi-shot configurations yield superior Macro F1 scores, holds true across all languages, as Table 4 shows, confirming their ability to provide a more faithful assessment of true classification capabilities.

Language	Strategy	Accuracy	Macro-F1	Micro-F1
Chinese	0 shot	87.50%	85.39%	87.50%
	1 shot	91.00%	89.78%	91.00%
	3 shot	**91.50%**	**90.39%**	**91.50%**
	5 shot	91.00%	89.78%	91.00%
	5 shot_COT	91.00%	89.74%	91.00%
Japanese	**0 shot**	**92.25%**	71.64%	**92.25%**
	1 shot	89.00%	69.44%	89.00%
	3 shot	80.75%	63.01%	80.75%
	5 shot	91.75%	**73.12%**	91.75%
	5 shot_COT	89.50%	68.73%	89.50%
English	0 shot	75.50%	69.78%	75.50%
	1 shot	75.00%	68.83%	75.00%
	3 shot	77.25%	71.23%	77.25%
	5 shot	**78.00%**	**72.26%**	**78.00%**
	5 shot_COT	75.75%	71.84%	75.75%

Table 4: Promise Status Identification Performance by Language and Strategy.

Language	Strategy	Accuracy	Macro-F1	Micro-F1
Chinese	0 shot	82.25%	55.98%	82.25%
	1 shot	87.00%	73.79%	87.00%
	3 shot	87.75%	75.81%	87.75%
	5 shot	**88.00%**	**76.14%**	**88.00%**
	5 shot_COT	84.75%	69.89%	84.75%
Japanese	**0 shot**	**69.75%**	43.80%	**69.75%**
	1 shot	69.25%	44.60%	69.25%
	3 shot	69.00%	**45.43%**	69.00%
	5 shot	69.25%	44.94%	69.25%
	5 shot_COT	68.50%	44.61%	68.50%
English	0 shot	73.50%	72.80%	73.50%
	1 shot	74.75%	74.33%	74.75%
	3 shot	**75.25%**	**75.01%**	**75.25%**
	5 shot	72.75%	72.22%	72.75%
	5 shot_COT	73.50%	72.56%	73.50%

Table 5: Evidence Status Assessment Performance by Language and Strategy.

4.2.2 Evidence Status Assessment

Assessing Actionable Evidence requires relational reasoning and is substantially more complex than PI. Table 5 shows Chinese performance scaling with context, improving from 82.25% in the zero-shot setting to a peak of 88.00% in the 5-shot setting. In contrast, Table 5 also shows Japanese performance plateauing at 69.75% in the zero-shot setting and failing to improve with demonstrations. This stagnation suggests a high sensitivity to linguistic nuance that general LLM prompts struggle to capture. This observation is supported by the SemEval findings (Chen et al., 2025), where the WC Team achieved strong performance in Japanese evidence identification by utilizing the language-specific Tohoku-BERT model, suggesting that capturing language-specific writing styles is critical for evidence evaluation.

4.2.3 Evidence Quality Evaluation

Table 6 presents Evidence Quality Evaluation as the most challenging subtask, yielding the lowest and most variable Macro-F1 scores. Table 6 further shows the highest Macro-F1 in Chinese (40.98%) under the 5-shot + CoT strategy. This result confirms that explicit reasoning provides a marginal benefit in this fine-grained judgment task. The inferential difficulty of EQE is directly related to the assessment of misalignment. Table 1 indicates that misleading cases are rare in the full dataset—1 in Chinese and 23 in Japanese—yet they pose a significant risk and often involve superficial evidence or the linking of unrelated past data to future policies.

4.2.4 Verification Timeline Prediction

Verification Timeline Prediction yielded highly language-dependent outcomes. Table 7 indicates Chinese performance peaking at 83.00% Accuracy in the 5-shot setting, aligning with the distribution in which Chinese samples skew toward short-term verification. By contrast, Table 7 reports that few-shot learning did not improve English performance, with the zero-shot baseline remaining highest at 49.25%. Table 1 documents a large share of English samples labeled “Other” (245), suggesting indefinite timelines or non-temporal constraints that are underrepresented in the demonstrations. The task is further complicated by large corporations balancing short-term verification with long-term goals extending beyond five years, as Table 4 highlights.

4.3 Best-Case Performance by Language

Aggregating the optimal configurations reveals a marked cross-lingual disparity. Table 8 reports Chinese with the highest average Accuracy at 85.12%, substantially exceeding Japanese at 68.94% and English at 63.62%. This apparent Chinese advantage chiefly reflects dataset characteristics and severe class imbalance: SemEval analyses (Chen et al., 2025) note that the Chinese subset—owing to annotation methodology—contains a much lower proportion of positive samples. Table 1 documents this pattern in Actionable Evidence, where Chinese samples are predominantly labeled “No” (832 in the full dataset). As a result, Accuracy and Macro-F1 diverge widely for Chinese—Table 6 shows Evidence Quality at 78.00% Accuracy versus 40.98% Macro-F1—underscoring the need to treat Macro-F1 as the primary, less biased metric for fair cross-lingual comparison.

Language	Strategy	Accuracy	Macro-F1	Micro-F1
Chinese	0 shot	76.25%	39.35%	76.25%
	1 shot	77.50%	**41.01%**	77.50%
	3 shot	77.00%	39.71%	77.00%
	5 shot	77.50%	39.97%	77.50%
	5 shot_COT	**78.00%**	40.98%	**78.00%**
Japanese	0 shot	32.75%	21.16%	32.75%
	1 shot	37.25%	23.17%	37.25%
	3 shot	**38.50%**	**23.49%**	**38.50%**
	5 shot	36.25%	23.20%	36.25%
	5 shot_COT	36.50%	22.86%	36.50%
English	0 shot	44.25%	32.81%	44.25%
	1 shot	43.75%	32.16%	43.75%
	3 shot	**52.00%**	**37.15%**	**52.00%**
	5 shot	46.50%	33.39%	46.50%
	5 shot_COT	50.25%	36.40%	50.25%

Table 6: Evidence Quality Evaluation Performance by Language and Strategy.

Language	Strategy	Accuracy	Macro-F1	Micro-F1
Chinese	0 shot	76.75%	30.27%	76.75%
	1 shot	79.00%	35.83%	79.00%
	3 shot	80.75%	45.40%	80.75%
	5 shot	**83.00%**	**48.36%**	**83.00%**
	5 shot_COT	81.00%	43.22%	81.00%
Japanese	0 shot	74.50%	28.90%	74.50%
	1 shot	**75.25%**	27.76%	**75.25%**
	3 shot	74.50%	31.71%	74.50%
	5 shot	73.25%	30.64%	73.25%
	5 shot_COT	70.25%	**32.22%**	70.25%
English	**0 shot**	**49.25%**	**20.04%**	**49.25%**
	1 shot	46.75%	18.71%	46.75%
	3 shot	45.50%	18.10%	45.50%
	5 shot	46.25%	18.98%	46.25%
	5 shot_COT	48.00%	19.45%	48.00%

Table 7: Verification Timeline Prediction Performance by Language and Strategy.

Task	Chinese	Japanese	English
Promise Status	91.50%	**92.25%**	78.00%
Evidence Status	**88.00%**	69.75%	75.25%
Evidence Quality	**78.00%**	38.50%	52.00%
Verification Timeline	**83.00%**	75.25%	49.25%
Average	**85.12%**	68.94%	63.62%

Table 8: Best Performance by Task and Language (Highest Accuracy Configuration).

Task	With COT(Accuracy)	Without COT(Accuracy)	With COT(Macro-F1)	Without COT(Macro-F1)
Promise Status	85.42%	**86.92%**	76.77%	**78.39%**
Evidence Status	66.42%	**67.50%**	31.63%	**32.66%**
Evidence Quality	75.58%	**76.67%**	62.35%	**64.44%**
Verification Timeline	**54.92%**	53.42%	**33.41%**	32.19%

Table 9: Effectiveness of CoT Reasoning (Accuracy, %)

4.4 Impact of Chain-of-Thought Reasoning

Table 9 shows that the utility of CoT reasoning is highly task-dependent: when averaged across all languages and tasks, CoT yields a small but consistent aggregate decline—Accuracy decreases by 0.54 pp and Macro-F1 decreases by 0.88 pp. This confirms that CoT introduces unproductive processing overhead for tasks driven by direct semantic pattern matching. A phenomenon consistent with recent findings that step-by-step reasoning can actively degrade model accuracy in tasks resembling human overthinking scenarios (Liu et al., 2024).

However, CoT proved selectively effective, providing a measurable benefit in the Evidence Quality Evaluation subtask. This utility is maximized in inferentially complex scenarios demanding explicit, structured reasoning—such as assessing the likelihood of greenwashing. Therefore, CoT should be reserved for nuanced alignment tasks where multi-step judgment is required, rather than being applied as a default strategy for generalized classification.

5 Conclusion

This study aims to establish a multilingual evaluation framework for ESG promise verification using Large Language Models (LLMs) and to assess model performance across four verification subtasks in Chinese, Japanese, and English sustainability reports.

Our findings indicate that few-shot prompting, particularly the 5-shot configuration, provides more stable and reliable classification outcomes than zero-shot prompting, while the relative task difficulty differs significantly. Promise Identification is comparatively more tractable, whereas Evidence Quality Evaluation requires more complex contextual reasoning and remains the most challenging. Chain-of-Thought reasoning is not universally beneficial but demonstrates selective improvements in nuanced inference tasks.

In terms of academic contribution, this study provides a systematic benchmark for multilingual ESG promise verification and clarifies how task complexity, linguistic variation, and annotation imbalance jointly influence model reasoning behavior. It also enriches understanding of prompt design effects in cross-lingual sustainability contexts.

Regarding managerial implications, our results offer actionable guidance for ESG governance stakeholders. Organizations seeking to reduce greenwashing risks may adopt AI-driven verification mechanisms to enhance sustainability disclosure transparency and consistency, while regulatory bodies and investors can leverage these tools to improve oversight, credibility assessment, and accountability in sustainable finance.

Future work may incorporate retrieval-augmented techniques or domain-specific model adaptation to improve evidence relevance, extend evaluation to additional languages and industries, and develop explainable reasoning outputs to support real-world audits and compliance processes in sustainability reporting.

Acknowledgments

This research was supported by the Industrial Technology Research Institute (ITRI) and National Taipei University (NTPU), Taiwan, under grants NTPU-114A513E01 and NTPU-113A513E01; the National Science and Technology Council (NSTC), Taiwan, under grant NSTC 114-2425-H-305-003-; and National Taipei University (NTPU) under grant 114-NTPU_ORDA-F-004.

References

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555-596. https://doi.org/10.1162/coli.07-034-R2

Chen, C.-C., Seki, Y., Shu, H., Lhuissier, A., Kang, J., Lee, H., Day, M.-Y., & Takamura, H. (2025, July). SemEval-2025 Task 6: Multinational, Multilingual, Multi-Industry Promise Verification. In S. Rosenthal, A. Rosá, D. Ghosh, & M. Zampieri, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)* Vienna, Austria.

Chen, X., Zeynali, A., Camargo, C., Flöck, F., Gaffney, D., Grabowicz, P., Hale, S. A., Jurgens, D., & Samory, M. (2022, July). SemEval-2022 Task 8: Multilingual news article similarity. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, & S. Ratan, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* Seattle, United States.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018, oct nov). XNLI: Evaluating Cross-lingual Sentence Representations. In E. Riloff, D. Chiang, J. Hockenmaier, & J. i. Tsujii, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* Brussels, Belgium.

Delmas, M. A., & Burbano, V. C. (2011). The Drivers of Greenwashing. *California Management Review*, *54*(1), 64-87. https://doi.org/10.1525/cmr.2011.54.1.64

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Lublóy, Á., Keresztúri, J. L., & Berlinger, E. (2025). Quantifying firm-level greenwashing: A systematic literature review. *Journal of Environmental Management*, *373*, 123399. https://doi.org/https://doi.org/10.1016/j.jenvman.2024.123399

Lyon, T. P., & Montgomery, A. W. (2015). The Means and End of Greenwash. *Organization & Environment*, *28*(2), 223-249. https://doi.org/10.1177/1086026615575332

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Seki, Y., Shu, H., Lhuissier, A., Lee, H., Kang, J., Day, M.-Y., & Chen, C.-C. (2024). ML-Promise: A Multilingual Dataset for Corporate Promise Verification. *arXiv preprint arXiv:2411.04473*.

Testa, F., Boiral, O., & Iraldo, F. (2018). Internalization of Environmental Practices and Institutional Complexity: Can Stakeholders Pressures Encourage Greenwashing? *Journal of Business Ethics*, *147*(2), 287-307. https://doi.org/10.1007/s10551-015-2960-2

Wang, X., Gao, X., & Sun, M. (2025). Construction and analysis of corporate greenwashing index: a deep learning approach. EPJ Data Science, 14(1), 44. https://doi.org/10.1140/epjds/s13688-025-00562-w

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024). Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. arXiv:2410.21333. https://doi.org/10.48550/arXiv.2410.21333

Exploring Sentence Stress Detection Leveraging Whisper-based Speech Foundation Models

Ting-An Hung*[1], Yu-Hsuan Hsieh*[1], Tien-Hong Lo[1], Yung-Chang Hsu[2], Berlin Chen[1]

[1]**National Taiwan Normal University,** [2]**EZAI Corp.**
{annhung323451, ivyhsieh0118}@gmail.com,
{teinhonglo, berlin}@ntnu.edu.tw, mic@ez-ai.com.tw

Abstract

Sentence stress reflects the relative prominence of words within a sentence. It is fundamental to speech intelligibility and naturalness, and is particularly important in second language (L2) learning. Accurate stress production facilitates effective communication and reduces misinterpretation. In this work, we investigate Sentence Stress Detection (SSD) using Whisper-based Transformer speech models under diverse settings, including model scaling, backbone–decoder interactions, architectural and regularization enhancements, and embedding visualization for interpretability. Results show that smaller Whisper variants outperform larger ones under limited data. With architectural and regularization enhancements, and by fixing a decoder whose capacity matches the dataset scale, both small and large backbones benefit. Consequently, even larger models can achieve competitive or superior performance under data-scarce conditions, partially mitigating data limitation effects. Embedding analysis reveals clear separation between stressed and unstressed words. These findings offer practical insights into model selection, architecture design, and interpretability for SSD applications, with implications for L2 learning support tools.

1 Introduction

Automatic detection of sentence stress in spoken language is crucial for speech intelligibility, prosodic naturalness, and perceived fluency, particularly in second language (L2) learning (Ladd, 2008; Lee et al., 2016; van Heuven, 2014). Misplaced stress in L2 learners can lead to misunderstandings and reduced comprehension, motivating the development of automated Sentence Stress Detection (SSD) systems for assessment and feedback (Lin et al., 2020; Kakouros and Räsänen, 2016).

Recent advances in pre-trained speech foundation models, such as Whisper, enable the extraction of rich embeddings that encode both acoustic and prosodic information (Radford et al., 2022; Bain et al., 2023). Whisper models, trained on massive multilingual corpora, can be adapted to downstream tasks like SSD without requiring extensive task-specific data (Nguyen et al., 2023; de Seyssel et al., 2023). Building on this, the WhiStress model (Yosha et al., 2025) demonstrated the effectiveness of Whisper embeddings for prosodic feature learning. However, systematic studies investigating how model size, architectural choices, and regularization strategies affect SSD performance and interpretability remain limited.

To this end, we explore Whisper-based SSD under diverse settings, including model scaling, backbone–decoder interactions, and architectural enhancements. We also analyze embedding representations for interpretability (Van Heuven, 2018; Arvaniti, 2020).

Our main contributions are as follows:

- evaluating Whisper-based SSD across multiple model sizes,
- analyzing the impact of decoder configuration, architectural enhancements, and regularization,
- conducting embedding visualization to interpret stress representations.

2 Related Work

2.1 Sentence Stress Detection

Early studies on SSD relied on handcrafted acoustic-prosodic features such as pitch (F0), intensity, and duration, modeled using Support Vec-

*Equal contribution.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 314–319
November 20-22, 2025 ©2025 Association for Computational Linguistics

tor Machines (SVMs) or Hidden Markov Models (HMMs).(Mishra et al., 2012; Auran et al., 2004). While effective in constrained settings, these methods required extensive domain expertise and failed to capture complex interactions among prosodic cues. Subsequent deep learning approaches, using Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Transformer architectures (Vaswani et al., 2017), advanced SSD by learning hierarchical features directly from raw speech (Baevski et al., 2020; Pasad et al., 2021). However, these models remain data-intensive and often lack interpretability, particularly in low-resource L2 contexts.

2.2 Pre-trained Speech Models and Whisper

Self-supervised models such as Wav2Vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) learn general-purpose speech representations that encode both phonetic and prosodic cues, facilitating transfer to downstream tasks. More recently, Whisper (Radford et al., 2022) introduced a large-scale encoder–decoder architecture trained on massive multilingual corpora. Whisper embeddings have demonstrated utility beyond ASR, including emotion recognition and prosodic analysis (Nguyen et al., 2023). Building on this foundation, the WhiStress model (Yosha et al., 2025) adapted Whisper for SSD using an alignment-free framework, but its evaluation was limited to a single variant (Whisper-small.en). Broader studies examining scaling behavior, architectural design, and regularization effects remain scarce.
Although pre-trained speech models have significantly advanced prosodic modeling, there has been limited investigation into how model size, architecture, and regularization influence SSD performance and interpretability.

3 Method

3.1 Model Architectures and Configurations

Our architecture, which is similar to WhiStress (Yosha et al., 2025), is based on the Whisper encoder to extract speech embeddings that implicitly encode acoustic and prosodic cues. A stress decoder then predicts word-level stress labels. The overall architecture is shown in Figure 1. Compared with WhiStress, which directly applies Whisper embeddings to a classification head, our design introduces several modifications: Backbone–decoder scaling:

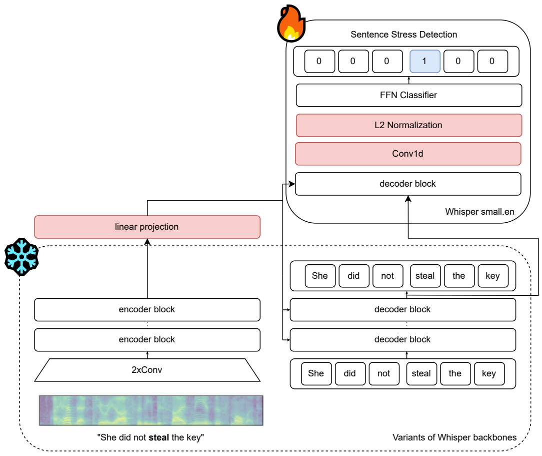


Figure 1: Overall architecture of the proposed SSD model. Input speech is processed by the Whisper encoder, followed by a fixed or trainable stress decoder. Optional components, including Conv1D, L2 normalization, and dropout, are applied depending on the model configuration.

varying both encoder size and decoder capacity. Following WhiStress (Yosha et al., 2025), which utilized the 9th layer out of 12 encoder layers in Whisper-small for optimal stress representation, intermediate-to-upper encoder layers were found to capture prosodic and phonetic cues most effectively. Accordingly, We generalize this approach by selecting approximately three-quarters of the encoder layers for each Whisper variant, namely 3, 5, 9, 18, and 24 layers for the tiny, base, small, medium, and large models, respectively.

This proportional selection strategy aims to preserve high-level prosodic features while maintaining training efficiency and mitigating overfitting risks.

We incorporate architectural enhancements by adding Conv1D and projection layers to better capture local prosodic dynamics. We also apply regularization mechanisms, including dropout and L2 normalization, to improve training stability and model generalization.

3.2 Experimental Configurations

Configuration I: Joint Scaling of Backbone and Stress Decoder. In this configuration, both the Whisper backbone and the stress decoder vary between Base, Small, Medium, and Large (Radford et al., 2022). This allows us to examine how the overall model capacity affects SSD, including potential overfitting for larger models.

Configuration II: Fixed Decoder, Varying Backbone. Here, the stress decoder is fixed as Whisper-Small.en (Radford et al., 2022), while the backbone varies in size. This isolates the contribution of the backbone to SSD performance, providing a fair comparison of different representation capacities without confounding changes in the classification head.

Configuration III: Architectural and Regularization Enhancements. The backbone is fixed at Whisper-Small.en. The stress decoder incorporates several enhancements:

- **Conv1D layer**: captures local temporal dependencies in frame-level embeddings, enhancing local prosodic pattern learning.
- **L2 normalization**: word-level embeddings x are normalized as $\hat{x} = x/\|x\|_2$, standardizing embedding magnitudes to improve generalization and stability.
- **Dropout**: randomly zeros out portions of embeddings during training to prevent overfitting, especially important for high-dimensional embeddings.

Configuration IV: Embedding Visualization and POS Analysis. We use t-SNE to project word-level embeddings and inspect stress clustering. Part-of-speech (POS) analysis examines whether specific word types, such as nouns, verbs, or function words, are more challenging, providing insight into systematic error patterns.

4 Experiments

4.1 Dataset

All experiments are conducted on the TINYSTRESS-15K dataset (Eldan and Li, 2023), a fully synthetic English speech corpus designed for SSD evaluation. It contains 15,000 training samples and 1,000 test samples, totaling approximately 15 hours of audio. Word-level stress annotations and precise time alignment are provided. Prosodic parameters such as pitch, duration, and amplitude are manipulated to simulate natural sentence stress, while multiple synthetic speaker voices increase variability. This controlled synthetic design allows for consistent evaluation of model performance under diverse prosodic variations without the need for costly manual labeling. However, as the dataset is fully synthetic, it may not perfectly capture the acoustic nuances of natural speech. Future work will include testing on natural speech corpora to further validate model performance.

4.2 Training Details

All experiments are trained for 20 epochs with batch size 16, using the AdamW optimizer (Kingma and Ba, 2015) with an initial learning rate of 1e$-$4 and cross-entropy loss. Whisper backbones are frozen, and only the stress decoder is updated. The training set is split into 90% for training and 10% for validation. No early stopping is applied; the model achieving the best F1 score on the validation set among all 20 epochs is used for reporting results. Models are evaluated using word-level F1 score, precision, and recall.

4.3 Results

Table 1 summarizes the performance of different backbone-decoder combinations. To isolate the effect of backbone representation power, we fix the decoder as Whisper-Small.en and vary the backbone size, as shown in Table 2.

Model	Precision	Recall	F1
Tiny	0.8733	0.8576	0.8653
Base	0.8834	0.8885	0.8859
Small	**0.9301**	**0.9288**	**0.9294**
Medium	0.8399	0.8381	0.8390
Large	0.7309	0.8245	0.7748

Table 1: Configuration I: Joint scaling of backbone and stress decoder. Larger models do not necessarily improve performance, likely due to overfitting.

Backbone	Precision	Recall	F1
Tiny	0.8664	0.8957	0.8808
Base	0.8726	0.9065	0.8892
Small	0.9348	0.9187	0.9267
Medium	**0.9509**	**0.9612**	**0.9560**

Table 2: Configuration II: Fixed decoder (Small.en) with varying backbone sizes. Small backbone provides optimal trade-off between capacity and generalization, while Medium achieves the highest F1.

5 Discussion

Configuration I: Joint Scaling Larger models do not consistently improve SSD performance; the Large backbone shows overfitting under limited data. Small and Base achieve better generalization, suggesting that model capacity must be

Enhancement	Precision	Recall	F1
Baseline	0.8664	0.8957	0.8808
Dropout only	0.9095	0.9324	0.9208
L2 normalization only	0.9125	0.8928	0.9025
Conv1D only	0.9209	0.9302	0.9256
Conv1D + L2 normalization	0.9366	0.9144	0.9254
Dropout + L2 normalization	0.8941	0.9295	0.9115
Dropout + Conv1D	**0.9364**	**0.9317**	**0.9340**

Table 3: Configuration III: Ablation study of architectural and regularization enhancements on SSD performance using the Whisper-Tiny backbone with a fixed Small decoder.

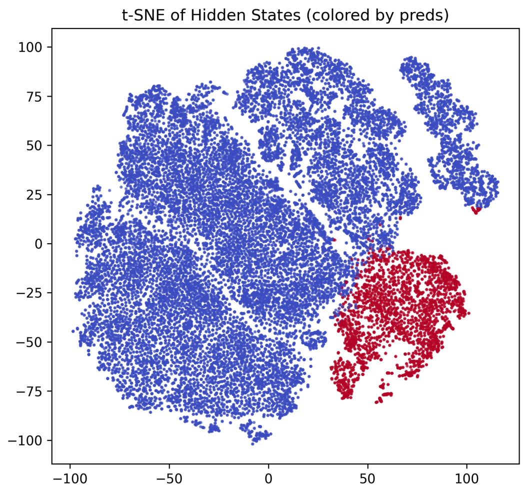


Figure 2: Configuration IV: t-SNE visualization of word-level embeddings. Red points = stressed words, blue = unstressed. POS analysis indicates nouns and verbs cluster more distinctly than function words, suggesting systematic differences in classification difficulty.

matched with data scale. The lack of linear projection may further limit larger models, as seen in the improvements from architectural enhancements in Configuration III.

Configuration II: Fixed Decoder, Varying Backbone With the decoder fixed, the medium backbone achieves the best F1 (0.9560), confirming that the backbone size directly impacts SSD quality. Small still balances accuracy and efficiency, making it practical in resource-constrained settings.

Configuration III: Architectural Enhancements Ablation in the tiny backbone shows that Conv1D (F1 = 0.9256), Dropout (0.9208) and L2 normalization (0.9025) each improve performance over baseline (0.8808). Additional analysesexamining partial module combinations reveal that: using only Dropout + L2 normalization yields F1 = 0.9115; only Conv1D + L2 normalization gives F1 = 0.9253; and only Conv1D + Dropout results in F1 = 0.9340, making it the bbest performingpartial module combination. Gains on larger backbones were marginal, indicating current performance may be bounded by dataset size. These targeted enhancements remain crucial for stable training on limited data.

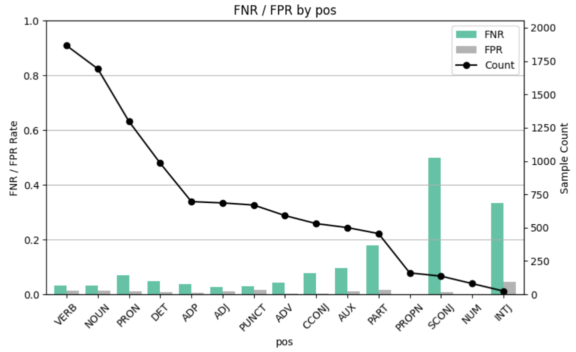


Figure 3: FNR/FPR by POS with sample counts (line). High-frequency POS (VERB/NOUN) show low FNR and near-zero FPR, while low-frequency categories such as SCONJ and NUM exhibit higher FNR despite smaller counts.

Configuration IV: Embedding and POS Analysis t-SNE visualizations of word-level embeddings show a clear separation between stressed and unstressed words, with stressed words forming more compact clusters. POS analysis further reveals that function words are more error-prone compared to content words. In a more detailed breakdown, high-frequency categories such as VERB, NOUN, PRON, DET, and ADP exhibit low FNR and near-zero FPR, indicating reliable predictions, while low-frequency categories like SCONJ, NUM, and INTJ have high FNR but low FPR, meaning many true instances are missed but mislabeling is rare. These observations suggest that improving VERB prediction and increasing data for rare POS, as well as incorporating POS-aware modeling or additional prosodic cues, could enhance the overall performance.

6 Conclusion and Future Work

This study systematically investigated Sentence Stress Detection (SSD) using Whisper-based models, focusing on model scaling, decoder configuration, architectural enhancements, and embedding interpretability. Results show that scaling backbone and decoder simultaneously may cause overfitting under limited data, while fixing the decoder provides a clearer evaluation of backbone capacity, benefiting both small and large backbones. Consequently, larger models can achieve competitive or superior performance under data-scarce conditions, partially mitigating data limitation effects. Lightweight modifications such as Conv1D, L2 normalization, and dropout improve robustness, and embedding analyses reveal both stress separability and systematic misclassification patterns, particularly for smaller backbones. POS analysis indicates that function words are more challenging, suggesting potential benefits from POS-aware modeling or additional prosodic cues.

Future work will extend SSD to multilingual and cross-lingual contexts, incorporate richer linguistic features (e.g., syllable structure, phonological rules, POS embeddings), and evaluate real-world scenarios including noisy, spontaneous, and accented speech. Multi-layer embedding fusion may further capture complementary prosodic cues. Finally, given current performance appears constrained by dataset scale, exploring data-efficient strategies such as semi-supervised learning, augmentation, or active learning will be critical to overcoming data scarcity and advancing SSD performance.

Overall, this work provides practical insights for designing SSD models and informs the development of L2 learning support tools, offering both quantitative and qualitative guidance for future research.

References

Amalia Arvaniti. 2020. The phonetics of prosody.

Cyril Auran, Caroline Bouzon, and Daniel J. Hirst. 2004. The aix-marsec project: an evolutive database of spoken british english. *Speech Prosody 2004*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

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

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english?

Vincent J. van Heuven. 2014. Acoustic correlates and perceptual cues of word and sentence stress: Mainly english and dutch. In *INTERSPEECH 2014*.

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 Transactions on Audio, Speech, and Language Processing*.

Sofoklis Kakouros and Okko Räsänen. 2016. 3pro – an unsupervised method for the automatic detection of sentence prominence in speech. *Speech Communication*, 82:67–84.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization.

D. Robert Ladd. 2008. *Intonational Phonology*, 2 edition. Cambridge Studies in Linguistics. Cambridge University Press.

Gary Lee, Ho-Young Lee, Jieun Song, Byeongchang Kim, Sechun Kang, Jinsik Lee, and Hyosung Hwang. 2016. Automatic sentence stress feedback for non-native english learners. *Computer Speech & Language*, 41.

Binghuai Lin, Liyuan Wang, Xiaoli Feng, and Jinsong Zhang. 2020. Joint detection of sentence stress and phrase boundary for prosody. In *INTERSPEECH*, pages 4392–4396.

Taniya Mishra, Vivek Rangarajan Sridhar, and Alistair Conkie. 2012. Word prominence detection using robust yet simple prosodic features. In *Interspeech 2012*, pages 1864–1867.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, and Michael Hassid. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Maureen de Seyssel, Antony D'Avirro, Adina Williams, and Emmanuel Dupoux. 2023. Emphassess: a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. *arXiv preprint arXiv:2312.14069*.

Vincent Van Heuven. 2018. *Acoustic Correlates and Perceptual Cues of Word and Sentence Stress: Theories, Methods and Data*, pages 15–59.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Iddo Yosha, Dorin Shteyman, and Yossi Adi. 2025. Whistress: Enriching transcriptions with sentence stress detection. In *Interspeech 2025*.

融合序列訊息與圖結構之反洗錢異常行爲分析
Integrating Sequential Information and Graph Structures for Anti-Money Laundering Anomaly Detection

Yin-Ju Wu[1], Gavin Tseng[2], Berlin Chen[1]
[1]National Taiwan Normal University
[2]VIS Experimental High School
[1]{61247075s, berlin}@ntnu.edu.tw
[2]cheetahbooked@gmail.com

摘要

反洗錢 (Anti-Money Laundering, AML) 是金融科技領域的重要研究課題，其目標在於識別潛在的可疑帳户與交易。然而隨著跨境支付與新型態交易的興起，洗錢行爲往往具有高度隱匿性與複雜的網路結構，傳統規則式方法在偵測效能與泛化能力上皆表現不足。近年來，雖然有研究嘗試將機器學習或深度學習方法應用於 AML，但仍存在許多挑戰。爲了解決這些問題，本研究提出一個基於序列圖融合的 AML 帳户風險預測框架。該方法的核心在於同時建模帳户的個體時序行爲與其在交易網路中的結構特徵。首先，將每個帳户的交易歷史分解爲入邊和出邊序列，使用雙分支 GRU 架構分別編碼，捕捉帳户的時序交易模式，接著使用雙向注意力圖卷積層，通過差異感知的消息傳遞機制同時處理正向和反向鄰居關係，學習帳户間的行爲差異，並通過注意力機制自適應融合節點自身特徵與雙向鄰居聚合特徵。此外，針對 AML 資料集的極度不平衡特性，引入類別重加權與平衡採樣策略。我們在公開的反洗錢資料集上驗證所提方法，實驗結果顯示該框架在極度不平衡的情境下能取得穩定的 F1 表現，相較於傳統基線方法具有顯著優勢。

Abstract

Anti-Money Laundering (AML) is a critical research area in Financial Technology (FinTech) focused on detecting suspicious financial activity. However, the rise of new transaction types has led to increasingly subtle and complex money laundering schemes, rendering traditional rule-based methods inadequate for both detection and generalization. While machine learning and deep learning offer a promising alternative, there are still many challenges. To address these challenges, we propose an AML prediction framework based on sequence-graph fusion. Its core innovation is the joint modeling of an account's individual temporal behavior and its structural features within the transaction network. Our approach begins by decomposing each account's transaction history into incoming and outgoing sequences, which are encoded via a dual-branch Gated Recurrent Unit (GRU) to capture nuanced temporal patterns. We then utilize a bidirectional attention-based graph convolutional layer that employs a difference-aware message-passing mechanism to process relationships in both forward and backward directions, learning the behavioral contrasts between connected accounts. Through the attention mechanism, the model adaptively fuses each node's intrinsic features with the aggregated features from both forward and backward neighbors. To counteract the extreme class imbalance inherent in AML data, our framework incorporates class re-weighting and balanced sampling strategies. We validated our proposed method on a public AML dataset. The experimental results demonstrate that our approach achieves stable F1-scores under severely imbalanced datasets, significantly outperforming traditional baseline methods.

關鍵字：反洗錢 (AML)；圖卷積網路；金融詐欺偵測

Keywords: Anti-Money Laundering (AML); Graph Convolutional Network; Financial Fraud Detection

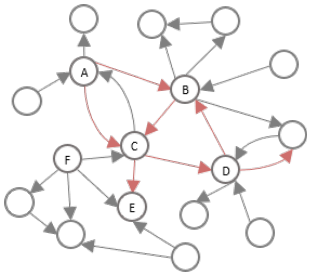
1 前言

在全球化金融體系快速發展的今日，反洗錢（Anti-Money Laundering, AML）已成爲金融監管與資訊科技研究的核心議題之一。根據國際洗錢防制組織（Financial Action Task Force, FATF）發布的指引與統計報告顯示洗錢活動對全球金融體系和經濟造成重大威脅 (Financial Action Task Force, 2010)，並且會影響國

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 320–325
November 20-22, 2025 ©2025 Association for Computational Linguistics

Time-stamp	From Bank	Account	To Bank	Account	Amount Received	Receiving Currency	Amount Paid	Payment Currency	Payment Format	Is Laundering
2022/9/1 12:20:00	1	A	1	B	1000	USD	1000	USD	Credit Card	1
2022/9/1 12:22:00	1	A	3	C	2500	USD	2500	USD	Credit Card	1
2022/9/1 12:35:00	1	B	3	C	1200	USD	1200	USD	Credit Card	1
					⋮					
2022/9/7 16:31:17	2	D	1	B	900	USD	900	USD	ACH	0

(a) 原始交易資料結構



(b) 轉換後的圖結構

Figure 1: 表格數據到圖結構的轉換

家的聲譽和金融系統的穩定性，更與跨國犯罪、毒品走私等高度相關。傳統的 AML 系統多依賴規則式方法 (Chen et al., 2018)，例如金額閾值檢測、異常交易模式識別與黑名單比對。然而，這些方法在面對日益複雜的交易網路與高隱匿性的洗錢手法時，往往面臨準確率不足與誤報率偏高的問題。

近年來，圖神經網路（Graph Neural Networks, GNNs）的崛起爲 AML 提供了新的解決思路 (Weber et al., 2019)(Johannessen and Jullum, 2023)。金融交易可以建模爲圖結構：帳户可視爲節點，交易可視爲邊，而交易金額、幣別與時間戳等屬性則成爲邊的特徵。透過圖建模方法，我們能捕捉到帳户間的交互關係與交易模式，並利用 GNN 進行節點分類，進而識別潛在的高風險帳户。這種方法具有兩大優勢：能從資料中自動學習複雜的異常行爲特徵，而不僅依靠人工設計的規則；能透過鄰居關係與結構訊息，偵測到更隱晦的洗錢模式，例如環狀交易或跨境分散匯款。

然而，單純的圖結構訊息仍不足以涵蓋洗錢行爲的全部特徵 (Tariq and Hassani, 2025)。許多可疑帳户的異常往往表現在交易時間序列上，例如短時間內進行多筆分散轉帳，或在固定週期內重複出現可疑資金流。若忽略時間維度，僅依靠靜態圖特徵，很可能無法正確區分正常高頻使用者與眞正從事非法活動的帳户 (Ghimire, 2023)。爲了同時捕捉交易的時序特徵與網路結構特徵，序列圖融合的建模思路逐漸受到關注 (Egressy et al., 2024)。這類方法最早在加密貨幣異常檢測領域得到成功應用 (Ding et al., 2024)，展現出結合時序行爲分析與圖結構學習的優勢。受此啓發，我們提出了一個專門針對 AML 任務的序列與圖融合框架，系統性地整合帳户時序行爲建模與交易網路結構分析，以實現更精確的異常帳户識別。

首先，我們將其交易歷史按方向性分解爲入邊序列和出邊序列，分別使用 GRU 編碼器學習其隱含的時序模式。這種設計能夠區分帳户的「資金流入」和「資金流出」的行爲特徵，對於識別不同類型的洗錢角色具有重要意義。其次，我們提出了雙向圖卷積層，專門處理有向交易圖中的複雜鄰居關係。與傳統圖卷積僅考慮無向鄰居不同 (Kipf and Welling, 2017)，我們的方法同時聚合正向鄰居（帳户指向的節點）和反向鄰居（指向該帳户的節點）的資訊 (Rossi et al., 2023)，並採用差異感知的消息函數，同時編碼節點間的行爲差異和鄰居原始特徵。這種設計使模型能夠學習到帳户間行爲模式的相似性和差異性 (Pahng and Hormoz, 2025)，進而識別異常的交易關係。

爲了解決 AML 資料中普遍存在的極度類別不平衡問題，我們採用了多重策略：包括基於類別頻率的加權損失函數、過採樣技術 (Chawla et al., 2002) 以及平衡採樣策略，確保模型能夠有效學習稀有的異常模式。本研究使用公開的大規模模擬 AML 資料集進行實驗驗證。該資料集專爲研究反洗錢與交易異常偵測而設計，包含多種已標註的洗錢模式，能夠模擬眞實金融網路中的轉帳行爲。實驗結果表明，我們提出的序列圖融合框架在多個評估指標上均優於傳統基線方法，特別是在極度不平衡的測試場景下仍能保持穩定的性能表現。

綜合而言，本文提出包括：(1) 一個序列與圖融合框架，在 AML 場景中系統性地結合了帳户時序行爲建模與網路結構分析；(2) 設計了雙向圖卷積層與差異感知消息傳遞機制，有效捕捉有向金融網路中的複雜鄰居關係；(3) 引入自適應特徵融合策略，通過注意力機制動態整合多源特徵。

2 方法

本研究採用圖結構的方法對金融交易網路進行建模，將複雜的資金流動關係轉換爲圖結構，以便後續的深度學習分析。

2.1 資料處理

原始 AML 資料集以關聯式表格形式儲存，每筆交易記錄包含時間戳、來源帳户、目標帳户、交易金額、幣別等基本屬性。爲了充分利用交易網路的拓撲特性，我們將表格數據 (如 Figure 1(a)) 轉換爲有向多重圖 (如 Figure 1(b)) 表示：帳户作爲節點，交易作爲邊，交易屬性作爲邊特徵。

考量到反洗錢研究的核心挑戰之一是避免未來資訊洩漏，我們在資料集的切割上採取時間導向分割的策略，而非隨機分割。遵循原始資料集的建議 (Altman, 2019)，我們將資料按 60%/20%/20% 的比例分割爲訓練集、驗證集和測試集，確保分割過程遵循時間順序原則。這種分割方式不僅符合實際金融監控場景的時序性要求，也確保了實驗結果的可重現性和實用性。

2.2 模型架構

我們的學習框架採用序列和圖融合的設計理念 (如 Figure 2)。對於交易圖中的每個帳户節點 v_i，我們首先將節點 i 的交易行爲分解爲入邊序列 S_i^{in} 和出邊序列 S_i^{out}：

$$S_i^{in} = \{t_1^{in}, t_2^{in}, ..., t_{m_i}^{in}\} \quad (1)$$
$$S_i^{out} = \{t_1^{out}, t_2^{out}, ..., t_{n_i}^{out}\} \quad (2)$$

其中每個交易 t_j 包含時間戳、金額、幣別等多維特徵。序列編碼採用雙分支 GRU 架構，分別處理入邊和出邊序列：

$$h_i^{in} = \text{GRU}_{in}(S_i^{in}) \quad (3)$$
$$h_i^{out} = \text{GRU}_{out}(S_i^{out}) \quad (4)$$

根據聚合策略，提取序列的最終表示：

$$x_i = \text{Concat}(h_i^{in}[-1], h_i^{out}[-1]) \quad (5)$$

此設計使得每個節點的特徵包含了該帳户完整的時序交易行爲資訊。

架構中使用的雙向圖卷積層能夠同時處理有向交易圖中的正向和反向資訊流。首先對所有節點的序列編碼特徵進行線性變換，接著對於每個節點 v_i，通過兩個獨立的消息傳遞過程聚合鄰居的序列編碼特徵：

$$h_i^{fwd} = \text{Propagate}(E, X) \quad (6)$$
$$h_i^{bwd} = \text{Propagate}(E^{-1}, X) \quad (7)$$

其中 $X = \{x_1, x_2, ..., x_N\}$ 爲所有節點的序列編碼特徵矩陣，E 爲原始有向邊集合（對應於帳户的出邊鄰居關係），E^{-1} 則爲反向邊集合（對應於帳户的入邊鄰居關係）。這種雙向傳播機制能夠分別聚合來自不同方向鄰居的行爲資訊。訊息傳遞的核心在於差異感知的訊息函數：

$$\text{m}(x_i, x_j) = \text{Linear}(\text{Concat}(x_i - x_j, x_j)) \quad (8)$$

這種訊息函數同時捕捉節點間的差異特徵（$x_i - x_j$）和鄰居原始特徵（x_j），使模型能夠學習到節點之間行爲模式的相似性和差異性。在每個圖卷積層內部，節點表示通過注意力機制融合三種特徵，自適應地權衡當前節點特徵與正向、反向鄰居特徵的重要性：

$$z_i = \text{Attention}([x_i, h_i^{fwd}, h_i^{bwd}]) \quad (9)$$

我們堆疊兩層這樣的雙向注意力圖卷積層以捕捉多跳鄰居關係，每層之間使用 ReLU 激活、批量正規化和 Dropout 正則，最終通過多層感知器解碼器進行二元分類。

我們的方法通過序列增強的圖表示學習，使圖卷積能夠處理包含完整交易歷史的豐富序列編碼。差異感知的消息傳遞機制同時編碼節點間的行爲差異和鄰居資訊，增強了異常模式的識別能力。雙向資訊聚合分別處理正向和反向邊，有效捕捉有向金融網絡的方向性特徵。此外，自適應特徵融合通過注意力機制動態權衡自身特徵與不同方向鄰居特徵的重要性。這種設計使得模型能夠同時利用帳户的個體時序行爲特徵和網絡中鄰居的行爲模式差異，實現對複雜洗錢行爲的精準識別。

3 實驗設置

3.1 資料集

我們採用了 (Altman et al., 2023) 開發的大規模合成 AML 資料集。該資料集通過虛擬世界中的個人、公司和銀行互動模型，生成了包含多種已標註洗錢模式的真實交易資料，能夠模擬眞實金融網路中的轉帳行爲。資料集包含兩種版本：HI 與 LI。前者的非法交易比例較高，適合用於訓練與初步驗證；後者的非法比例較低，更貼近眞實銀行場景中高度不平衡的分布特性。

在本研究中，我們聚焦於 Small 資料集進行實驗驗證。如 Table 1 所示，Small 子集包含 HI-Small（515K 帳户，5M 筆交易）和 LI-Small（705K 帳户，7M 筆交易），涵蓋時間跨度爲 10 天。這一規模既能有效檢驗模型

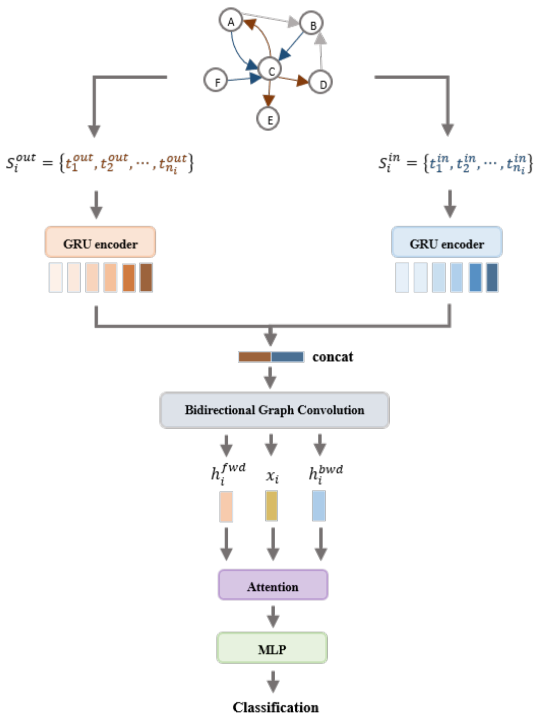


Figure 2: 模型架構圖

的分類能力，又能在有限的計算資源下保證多次重複實驗的可行性，特別適合於需要大量實驗組合的研究情境。

此外，Small 資料集同樣展現了典型的類別不平衡特徵：HI-Small 的非法交易比例約爲 1:981，而 LI-Small 的不平衡程度更爲嚴重，比例高達 1:1942。這種極端不平衡的分布充分反映了眞實金融場景中的 AML 問題，爲評估模型在實務應用中的表現提供了具有挑戰性且可信的測試基礎。

Table 1: AML 合成資料集（Small 子集）的詳細資訊

	Small	
Dataset Variant	**HI**	**LI**
Time Period (2022)	Sep 1–10	
Account Volume	515K	705K
Transaction Volume	5M	7M
Illicit Transactions	5.1K	4.0K
Fraud Ratio	1:981	1:1942

3.2 實驗設定

我們的實驗以交易記錄爲原始輸入，通過資料預處理構建交易圖結構並提取帳户交易序列。序列編碼部分採用 Gated Recurrent Unit（GRU），隱藏層維度 128，分別處理入邊與出邊序列並進行拼接。圖結構部分則使用兩層自定義雙向圖卷積層，整合注意力機制來自適應聚合鄰居訊息，每層輸出 128 維嵌入。序列與圖嵌入融合後，經由多層感知器輸出，最終完成二元分類。在訓練策略上，考量到類別高度不平衡，我們使用加權二元交叉熵損失（Weighted Binary Cross Entropy Loss）函數，根據類別頻率設定反比權重。採用 Adam 優化器，學習率設爲 1e-3，搭配 Early Stopping 策略與梯度裁剪確保訓練穩定性。模型選擇以驗證集 F1 分數爲準，當達到最佳驗證效果時保存模型參數。在評估上，我們採用 F1-score 作爲主要指標，並關注模型在極度不平衡資料下的整體表現，以評估其在眞實 AML 場景中的實用性。

3.3 評估指標

由於反洗錢任務具有高度不平衡的特性，我們採用多種衡量指標來全面評估模型表現。考量到實際應用中對檢測準確性與完整性的雙重要求，我們以 F1-score 作爲主要評估指標，其爲精確率（Precision）與召回率（Recall）的調和平均，能夠平衡兩者的重要性。其中，精確率衡量模型識別非法帳户的正確性，召回率則反映模型捕捉非法帳户的完整性。計算方式如下：其中，TP 爲眞正例（True Positives），TN 爲眞負例（True Negatives），FP 爲假正例（False Positives），FN 爲假負例（False Negatives）。計算方式如下：

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

F1-score 特別適合評估不平衡資料集的模型性能，因爲它既考慮了模型的檢測精度，也確保了對稀有類別（洗錢帳户）的有效識別。在 AML 實務應用中，過高的假警報率（低精確率）會造成調查資源浪費，而過高的漏報率（低召回率）則可能讓眞正的洗錢活動逃脫監管，因此 F1-score 提供了這兩個關鍵需求間的最佳平衡點。

4 實驗結果

我們在 HI-Small 與 LI-Small 兩個資料集上的實驗結果如 Table 2 所示，其中 **ours** 代表我們所提出的整合式架構。整體而言，該方法在兩個資料集上皆展現出最佳的整體效能：在 HI-Small 上 F1 分數達 0.85，在 LI-Small 上則爲 0.65。此結果顯示，我們的模型在維持

Table 2: HI-Small 與 LI-Small 資料集上不同序列輸入設定的實驗結果

Setting	HI-Small			LI-Small		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
in-only	0.9674	0.6312	0.7639	0.6579	0.4464	0.5319
out-only	**1.0000**	0.5461	0.7064	**0.9412**	0.2857	0.4384
combined	0.9730	0.5106	0.6698	0.6667	0.1071	0.1846
ours	0.9646	**0.7730**	**0.8583**	0.7949	**0.5536**	**0.6526**

高 Precision 的同時，能顯著提升 Recall，整體 F1 分數明顯優於其他設定。這驗證了所提出架構能同時捕捉交易網路中的時間動態與結構關聯，並在高度不平衡的資料情境下仍保持穩健與準確的偵測能力。

在此基礎上，我們進一步分析了「序列方向性資訊」對洗錢偵測效能的影響。爲了探討不同方向資訊的貢獻，我們設計了三種簡化的輸入設定進行比較：**in-only** 僅使用帳户作爲收款方的入邊序列；**out-only** 僅使用帳户作爲付款方的出邊序列；而 **combined** 則將兩個方向的交易資料合併成單一序列輸入至同一個 GRU 模組，不再保留方向性。相較之下，我們的方法 **ours** 採用雙分支 GRU，分別建模入邊與出邊序列，並結合雙向圖卷積層以同時學習帳户間的結構依賴與資金流向。

比較結果顯示，僅使用入邊序列時，雖然 Precision 仍維持高水準，但因缺乏出邊方向的輔助資訊，Recall 明顯下降，使得 HI-Small 與 LI-Small 的 F1 分數分別降至 0.76 與 0.53。若僅使用出邊序列，F1 分數進一步下降至 0.70 與 0.43，顯示單向付款行爲的資訊不足以支撐有效判別。當將兩個方向合併而不區分方向時，雖然 Precision 維持良好，但因方向性訊息喪失，模型難以辨識帳户在資金流動中的角色差異，最終導致 Recall 顯著下降，F1 僅達 0.66 與 0.18。

綜合上述結果，可以確認序列方向性資訊在洗錢交易偵測中的重要性，更同時驗證了我們的整合式架構的必要性。我們的完整架構透過序列模組捕捉序列的方向訊息，並藉由圖模組保留帳户之間的結構依賴，使模型能同時利用兩種異質訊息來源而達到更佳的檢測效能。

5 結論

本研究提出一個結合序列與圖結構資訊的完整架構，並透過 AML 資料集驗證其效能。實驗結果顯示，序列方向性資訊對檢測任務至關重要，而整合式架構能在不平衡情境下展現更佳的穩健性與準確性。與僅依賴靜態圖特徵或單純序列的方法相比，所提架構同時捕捉時間動態與結構依賴，能更有效識別可疑交易。未來研究可進一步探討其在更大規模資料集上的擴展性，並結合額外的領域知識以提升模型的表現與解釋性。

References

Erik Altman. 2019. IBM transactions for anti money laundering (AML). Kaggle Dataset.

Erik Altman, Béni Egressy, Jovan Blanuša, and Kubilay Atasu. 2023. Realistic synthetic financial transactions for anti-money laundering models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321—357.

Zhiyuan Chen, Le Dinh Van Khoa, Ee Na Teoh, Amril Nazir, Ettikan Kandasamy Karuppiah, and Kim Sim Lam. 2018. Machine learning techniques for anti-money laundering (aml) solutions in suspicious transaction detection: a review. *Knowledge and Information Systems*, 57(2):245–285.

Zhihao Ding, Jieming Shi, Qing Li, and Jiannong Cao. 2024. Effective illicit account detection on large cryptocurrency multigraphs.

Béni Egressy, Luc von Niederhäusern, Jovan Blanusa, Erik Altman, Roger Wattenhofer, and Kubilay Atasu. 2024. Provably powerful graph neural networks for directed multigraphs.

Financial Action Task Force. 2010. *Global Money Laundering and Terrorist Financing Threat Assessment*. FATF/OECD, Paris.

Sushrut Ghimire. 2023. Timetrail: Unveiling financial fraud patterns through temporal correlation analysis.

Fredrik Johannessen and Martin Jullum. 2023. Finding money launderers using heterogeneous graph neural networks.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*.

Seong Ho Pahng and Sahand Hormoz. 2025. Improving graph neural networks by learning continuous edge directions.

Emanuele Rossi, Bertrand Charpentier, Francesco Di Giovanni, Fabrizio Frasca, Stephan Günnemann, and Michael Bronstein. 2023. Edge directionality improves learning on heterophilic graphs.

Haseeb Tariq and Marwan Hassani. 2025. *Topology-Agnostic Detection of Temporal Money Laundering Flows in Billion-Scale Transactions*, page 402—419. Springer Nature Switzerland.

Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I. Weidele, Claudio Bellei, Tom Robinson, and Charles E. Leiserson. 2019. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics.

A Multi-faceted Statistical Analysis for Logit-based Pronunciation Assessment
一種用於發音評估的 Logit 多面向統計分析法

Chieh-Ren Liao
National Taiwan Normal University,
Taipei, Taiwan
80747002s@ntnu.edu.tw

Berlin Chen
National Taiwan Normal University, Taipei,
Taiwan
berlin@ntnu.edu.tw

摘要

發音品質評估中的發音好壞度 (Goodness of Pronunciation, GOP) 分數，是電腦輔助語言學習的關鍵技術。近期的研究指出，直接使用聲學模型原始輸出 logits 來計算 GOP 分數，其表現優於傳統基於 $softmax$ 機率的方法，因為 logits 避免了機率飽和問題並保留了更豐富的區分性資訊。然而，現有的 logit-based 方法大多僅依賴最大值、均值或變異數等基本統計量，這忽略了在音素持續時間內，logit 序列更為複雜的動態分佈與時序特性。為了更全面地捕捉 logit 序列中所蘊含的發音細節，本研究提出了一套多面向的統計分析法。我們探索了五種能夠描述 logit 序列不同特性的高階統計指標：(1) 動差生成函數，用以計算分佈的偏度 ($skewness$) 與峰度 ($kurtosis$)；(2) 資訊理論，透過計算熵 (entropy) 來量化模型的不確定性；(3) 高斯混合模型 (GMM)，用以擬合 logit 的多模態分佈；(4) 時間序列分析，計算自相關係數 ($autocorrelation$) 來衡量 logit 的穩定性；以及 (5) 極值理論，採用 top-k 平均來獲得更穩健的峰值信心度估計。我們在公開的 L2 英語語音資料庫 (SpeechOcean762) 上進行實驗，將這些新提出的統計指標與參考文獻中的基線方法 ($GOP_{MaxLogit}$, GOP_{margin}) 進行效能比較。初步結果顯示，部分高階統計指標，特別是那些能夠描述 logit 序列穩定性和分佈形狀的特徵，在發音錯誤檢測的分類任務上展現出更高的準確性，並與人類專家評分呈現出更強的相關性。這項研究證明，對 logit 序列進行更深層次的統計建模，是提升自動化發音評估系統效能的一個有效途徑。

Abstract

The Goodness of Pronunciation (GOP) score for pronunciation quality assessment is a key technology in computer-assisted language learning. Recent studies have shown that computing GOP scores directly from the acoustic model's raw output logits outperforms traditional softmax-probability-based methods, because logits avoid probability saturation issues and retain richer discriminative information. However, existing logit-based methods mostly rely on basic statistics such as maxima, means, or variances, which neglect the more complex dynamic distributions and temporal characteristics of logit sequences over phoneme durations. To more comprehensively capture pronunciation details embedded in logit sequences, this study proposes a multi-faceted statistical analysis method. We explore five higher-order statistical indicators that describe different characteristics of logit sequences: (1) moment-generating functions to compute distribution skewness and kurtosis; (2) information theory, using entropy to quantify model uncertainty; (3) Gaussian mixture models (GMMs) to fit multimodal distributions of logits;

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 326–333
November 20-22, 2025

(4) time-series analysis, computing autocorrelation coefficients to measure logit stability; and (5) extreme value theory, using top-k averaging to obtain more robust peak-confidence estimates. We conduct experiments on the public L2 English speech corpus SpeechOcean762, comparing these newly proposed statistical indicators with baseline methods from the literature ($GOP\_MaxLogit$, GOP_margin). Preliminary results show that some higher-order statistical indicators—particularly those that describe logit-sequence stability and distribution shape—achieve higher accuracy on pronunciation-error detection classification tasks and exhibit stronger correlation with human expert ratings. This study demonstrates that deeper statistical modeling of logit sequences is an effective approach to improving the performance of automated pronunciation assessment systems.

關鍵字：logit、gop

1 Introduction

在全球化時代，第二語言的口語溝通能力對於學術。然而，清晰的發音對 L2 學習者而言充滿挑戰，主要是因為母語 (L2) (L1) 的語音習慣會造成持續性的發音錯誤 。為此，電腦輔助發音訓練 (CAPT) 系統被廣泛發展，以提供即時且客觀的發音回饋 。在 CAPT 系統中，能夠在音素 (phoneme) 層級進行的發音錯誤檢測 (Mispronunciation Detection)，被證實對學習者改善特定發音問題特別有效 。

發音好壞度 (Goodness of Pronunciation, GOP) 是目前最主流的音素級別自動評估指標之一。傳統上，GOP 分數的計算依賴於深度神經網路 (DNN) 聲學模型輸出的後驗機率 (posterior probabilities) 。這些機率值是透過對模型的原始輸出 logits 進行 $softmax$ 歸一化得到的 。然而，$softmax$函數本身存在著「過度自信 (overconfidence)」的缺陷，容易將機率分佈推向極端，從而壓縮了不同音素之間的區分度，使得一些細微的發音偏差難以被偵測 。

為了解決 $softmax$ 歸一化的限制，Parikh et al. (2025)的研究開創性地提出直接使用未經處理的 logits 來計算 GOP 分數。相較於機率值，logits 保留了更豐富的鑑別資訊，並且避免了梯度飽和問題 。該研究探索了幾種基於 logit 的指標，例如最大 Logit ($GOP_{MaxLogit}$) ，用以捕捉模型的峰值信心 ；Logit 邊界 (GOP_{margin}) ，用以量化目標音素與其最主要競爭者之間的分離程度；以及 Logit 變異數 ($GOP_{VarLogit}$) ，用以衡量模型信心的穩定性。他們的實驗證明，在多數情況下，logit-based 的方法在發音錯誤檢測的分類任務上優於傳統的機率方法 。

儘管 Parikh et al. 的研究為 GOP 計算開闢了新的方向，但我們認為，他們所使用的方法仍有其侷限性。這些指標主要依賴 logit 序列的單點統計量 (如最大值) 或一階動差 (如均值、變異數)。這相當於將一個音素在持續時間內的 logit 變化視為一組無序的數字集合，忽略了其作為時間序列的內在結構以及其統計分佈的完整「形狀」。一個發音的過程是連續且動態的，其對應的 logit 序列在時間維度上的穩定性、對稱性與峰銳度，理應蘊含著關於發音品質的更深層線索。

基於此觀點，本研究旨在「超越均值與變異數」，提出一套更為全面且多面向的 logit 序列統計分析法。 我們不再僅僅滿足於 logit 的基本統計量，而是將其視為一個完整的統計分佈和時間序列來進行建模。我們系統性地引入了五類能夠從不同維度描述該序列特性的高階統計指標，包括：

- 分佈形狀特徵：透過計算偏度 ($skewness$) 與峰度 ($kurtosis$) 來捕捉 logit 分佈的不對稱性與集中趨勢。
- 資訊理論特徵：利用資訊熵 ($entropy$) 來量化模型在預測時的不確定性。
- 時序穩定性特徵：計算自相關係數 ($autocorrelation$) 來衡量 logit 序列隨時間變化的平滑程度。
- 分佈擬合特徵：採用高斯混合模型 (GMM) 來建模 logit 序列可能存在的多模態特性。
- 峰值穩健性特徵：透過極值理論中的 top-k 平均值來取代單一最大值，以獲得更可靠的峰值信心度。

我們將在公開的 SpeechOcean762 資料集 上驗證這些新指標的有效性，並與 Parikh et al. 的基線方法進行深入比較。本研究期望能證明，透過對 logit 序列進行更深層次的統計建模，我們能夠更精準地捕捉到發音的細微差異，從而為自動化發音評估技術開闢新的可能性。

2 研究方法

本研究旨在透過對 logit 序列進行更深層次的統計分析，來提升發音錯誤檢測的準確性。在本章節中，我們首先將簡要回顧作為我們比較基準的 logit-based GOP 指標。接著，我們將詳細闡述本研究提出的五類多面向統計特徵，這些特徵旨在從分佈形狀、資訊量、時間穩定性等多個維度，更全面地捕捉發音的細微動態。

2.1. 基線 $Logit - based\ GOP$ 指標 ($Baseline\ Logit - based\ GOP\ Metrics$)我們選用 Parikh et al. (2025) 所提出的主要 logit-based 指標作為效能比較的基線。這些指標代表了當前 logit-based GOP 方法的基礎。

最大 Logit ($GOP_{MaxLogit}$)：取音素對齊幀範圍內，目標音素 p 的 logit 序列 $l_t^{(p)}$ 中的最大值。此指標反映了模型在整個發音過程中所達到的最高信心水準。

$$GOP_{MaxLogit}(P) = \max_{t\in[t_1,t_2]} l_t^p \qquad \text{(式 1)}$$

Logit 邊界 (GOP_{margin})：計算在每一幀中，目標音素的 logit 值與最強競爭音素的 logit 值之間的差值，再將這些差值於整個音素段內取平均。此指標量化了目標音素在 logit 空間中的「突出程度」或「可區分性」。

$$GOP_{margin}(P) = \frac{1}{T}\sum_{t=t_1}^{t_2}\left(l_x^p - \max_{k\neq p} l_t^k\right) \qquad \text{(式 2)}$$

Logit 變異數 ($GOP_{VarLogit}$)：計算目標音素 logit 序列的變異數，用以衡量模型信心的穩定性。較低的變異數通常表示一個穩定、流暢的發音。

2.2. 提出的多面向統計指標 (Proposed Multi-faceted Statistical Metrics)

為了超越基線指標的侷限，我們引入了五類更為複雜的統計方法。這些方法被設計用來從 logit 序列中提取更深層次的資訊。

2.2.1. 分佈形狀特徵：動差分析 (Distribution Shape: Moment Analysis)

除了二階動差（變異數），更高階的動差能提供關於 logit 序列統計分佈「形狀」的額外資訊，這對於描述模型信心的動態變化至關重要。

- 偏度 ($Skewness, G1$)：作為第三階標準化動差，偏度衡量 logit 分佈的不對稱性。正偏度可能表示模型信心是逐漸建立然後迅速下降的過程，而負偏度則相反。異常的偏斜可能暗示著不自然的發音模式。
- 峰度 ($Kurtosis, G2$)：作為第四階標準化動差，峰度衡量分佈的「峰銳度」與「尾部

厚度」。高峰度表示模型的信心高度集中於某個值，伴隨可能的極端離群值；低峰度則表示分佈較為平坦。這有助於識別發音過程中信心的集中或分散程度。

2.2.2. 資訊理論特徵：不確定性量化 (Information Theory: Uncertainty Quantification)

此方法從整個後驗機率分佈的角度出發，而非僅僅關注目標音素，用以量化模型在預測時的整體「混淆程度」。

- 平均夏農熵 (Mean Shannon Entropy)：我們計算音素段內每一幀的後驗機率分佈x_t的夏農熵，然後取其平均值。熵是模型不確定性的直接度量。一個高的平均熵意味著模型的機率被分散在多個候選音素上，是發音含糊或錯誤的強烈信號。

$$H_{mean} = \frac{1}{T}\sum_{t=t_1}^{t_2} -\left(\sum_{k=1}^{D} P(k|x_t)\log P(k|x_t)\right)$$

(式 3)

- 平均 KL 散度 (Mean KL Divergence)：此指標衡量每一幀的實際後驗機率分佈 $P(x_t)$ 與一個代表「完美發音」的理想分佈（即目標音素機率為 1 的 one-hot 向量 Q）之間的「距離」。較大的 KL 散度意味著模型的輸出與理想狀態相去甚遠。

2.2.3. 分佈擬合特徵：高斯混合模型 (Distribution Fitting: Gaussian Mixture Models)

我們假設 logit 序列的分佈並非單峰，而是可能由多個潛在狀態（如音素的起始、穩定、結束階段）混合而成。高斯混合模型 (GMM) 能有效捕捉這種多模態特性。我們將 logit 序列擬合成一個包含 K 個高斯分量的 GMM，並提取其參數作為特徵，例如各分量的均值 (μ_k)、變異數 (σ_k^2) 和權重 (w_k)。這些參數能精細地描述發音過程中模型信心的多階段動態。

2.2.4. 時序穩定性特徵：自相關分析 (Temporal Stability: Autocorrelation Analysis)

為了彌補現有方法忽略 logit 序列時間順序性的不足，我們引入時間序列分析。我們計算 logit 序列在延遲為 1 (lag-1) 時的自相關係數 (Autocorrelation)。一個高的正相關係數表示 logit 序列是平滑且穩定變化的，這通常對應於一個清晰、穩定的發音。反之，一個接近於零或負值的係數則暗示著序列存在劇烈、不規則的波動，可能是發音不穩定的跡象。

2.2.5. 峰值穩健性特徵：極值理論 (Peak Robustness: Extreme Value Theory)

$GOP_{MaxLogit}$對單一的雜訊尖峰非常敏感。為了解決這個問題，我們採用一個更穩健的峰值估計方法：$top-k$ 平均值。此方法選取 logit 序列中最大的 k 個值（例如 k=3），並計算它們的平均值。這提供了一個更穩定的模型「峰值信心」的估計，有效地平滑了單一離群值的影響。

3. 實驗與結果 (Experiments & Results)

本章節將詳細闡述我們的實驗設計、評估指標，並對實驗結果進行深入的分析與討論。我們透過兩階段的實驗：初步實驗（使用 2500 筆訓練集）與完整實驗（使用 5000 筆完整資料集）：來全面評估我們提出的多面向統計指標的性能與穩定性。

3.1. 實驗設定 (Experimental Setup)

本研究的所有實驗皆於 SpeechOcean762 資料集上進行。此資料集提供了每個音素的標準音標與實際發音標註，讓我們得以客觀地產生「正確」與「錯誤」的標籤，作為分類任務的參考標準。

為了全面評估各項 GOP 指標的性能與穩定性，我們設計了兩組實驗：

1. 初步實驗：僅使用官方劃分的訓練集 (Training Split)，共 2500 筆語音，進行指標性能的初步探勘。

2. 完整實驗：使用完整資料集 (Full Dataset)，共 5000 筆語音，以驗證指標在更大、更多樣的數據上的泛化能力。

評估指標主要採用馬修斯相關係數 (MCC)，因其在處理類別不平衡數據時最具參考價值，同時也輔以準確率 (Accuracy)、精確率 (Precision)、召回率 (Recall) 與 $F1 - score$ 進行分析。

3.2. 初步實驗結果 (僅使用訓練集)

在僅使用 2500 筆訓練集數據的初步實驗中，我們發現描述 logit 分佈形狀的高階動差指標表現最佳。如表 1 所示，峰度 ($kurtosis$) 和偏度 ($skewness$) 在MCC分數上名列前茅，顯著優於傳統依賴 logit 數值大小的指標。然而，許多指標呈現出極端的「高召回、低精準」現象，顯示單一門檻值的分類能力有限。請見表 1。

發音錯誤檢測之分類效能比較 (以 MCC 分數排序)					
Method	Accuracy	Precision	Recall	F1-Score	MCC
kurtosis	0.661284	0.160985	0.422185	0.233090	0.081851
skewness	0.700848	0.163859	0.354305	0.224084	0.076690
gmm_weights_0	0.796730	0.186625	0.198675	0.192462	0.076373
gmm_vars_0	0.589221	0.145617	0.486755	0.224171	0.060064
gmm_vars_1	0.648365	0.146584	0.390728	0.213189	0.052300
autocorr_lag1	0.699435	0.149644	0.312914	0.202464	0.049562
logit_variance	0.183690	0.125924	0.958609	0.222607	0.043964
mean_logit_margin	0.129996	0.122761	0.998344	0.218637	0.027728
gmm_means_0	0.124142	0.122193	1.000000	0.217775	0.017578
gmm_weights_1	0.855268	0.144654	0.038079	0.060288	0.012652
evt_k3	0.124748	0.122114	0.998344	0.217611	0.010338
prosetrior_probability	0.144933	0.122453	0.975166	0.217584	0.009389
entropy_mean	0.122124	0.121946	1.000000	0.217383	0.005295
gmm_means_1	0.123738	0.121991	0.998344	0.217415	0.004471
max_logit	0.147961	0.122048	0.966887	0.216738	0.002055

表 1：在訓練集 (2500 筆) 上的分類效能

3.3. 完整實驗結果與分析 (使用完整資料集)

當我們將實驗擴展至全部 5000 筆數據時，結果發生了顯著且重要的變化。如表 2 所示，衡量「區分度」的 mean_logit_margin 以 MCC **0.3702** 的成績成為表現最佳的指標，而我們提出的「峰值信心」指標 evt_k3 和「模型確定性」指標 kl_to_onehot 緊隨其後。圖 1 中的橘色長條直觀地展示了這一點，一個由頂尖指標構成的「第一梯隊」已然成形。

如表 2 所示，基線方法中的 **mean_logit_margin** 以 **MCC 0.3702** 的成績位居榜首。這項結果極具說服力地證明，在數據充足的情況下，模型對於目標音素的判斷與其最主要競爭音素之間的「領先差距」，是判斷發音正確與否最為穩健和強大的單一指標。

發音錯誤檢測之分類效能比較 (以 MCC 分數排序)					
Method	Accuracy	Precision	Recall	F1-Score	MCC
mean_logit_margin	0.793070	0.176943	0.300600	0.222761	0.117813
kurtosis	0.854972	0.206271	0.165085	0.183394	0.105670
skewness	0.851258	0.201360	0.171195	0.185056	0.104221
gmm_weights_0	0.874960	0.232898	0.116639	0.155434	0.102638
prosetrior_probability	0.516339	0.127815	0.670158	0.214684	0.101243
entropy_mean	0.847082	0.194424	0.175014	0.184209	0.100268
gmm_means_1	0.651917	0.127739	0.433824	0.197364	0.069247
gmm_means_0	0.360902	0.110918	0.780906	0.194246	0.062045
max_logit	0.587120	0.120325	0.504746	0.194325	0.061078
evt_k3	0.782597	0.134983	0.222586	0.168053	0.053708
autocorr_lag1	0.878608	0.171588	0.060229	0.089162	0.046327
gmm_vars_1	0.857275	0.128065	0.076923	0.096115	0.024759
gmm_vars_0	0.132306	0.099640	0.970104	0.180719	0.016395
gmm_weights_1	0.885648	0.104607	0.021058	0.035059	0.002845
logit_variance	0.105967	0.098287	0.986361	0.178760	-0.012059

表 2：在完整資料集 (5000 筆) 上的分類效能

更值得注意的是，我們提出的兩種新方法: $\boldsymbol{evt\_k3}$ **(極值理論)** 和 $\boldsymbol{kl\_to\_onehot}$ **(KL 散度)** 以幾乎可以忽略不計的微小差距（MCC 分別為 0.3695 和 0.3685）緊隨其後。我們可以將這幾種表現最好的方法歸納為兩大類成功的策略：

1. **衡量「區分度與信心強度」的策略**：$mean\_logit\_margin$ 和 evt_k3 都屬於此類。前者量化了目標音素在 logit 空間中與其他音素的**分離程度**，而後者則以更穩健的方式（$top-k$平均）度量了模型信心的**峰值強度**。它們的成功表明，一個清晰、明確、無歧義的高信心分數，是正確發音最核心的聲學體現。
2. **衡量「模型總體確定性」的策略**：kl_to_onehot 和 $entropy\_mean$ 屬於此類。它們不只關心目標音素，而是評估整個輸出機率分佈的「混亂程度」。一個低的$entropy_mean$ 或 kl_to_onehot 值，代表模型對於其預測非常確定，幾乎沒有將機率分配給其他競爭者。這從另一個角度驗證了，模型的**整體判斷確定性**也是區分發音品質的關鍵。(請參見附錄一:表 2)

3.4. 比較分析與圖表解讀

圖 1 為我們的實驗發現提供了強而有力的視覺證明，我們可以從中得出三個核心的建設性結論：

1. **數據規模是性能的決定性因素**：從圖中可以一目了然地看到，幾乎所有指標的橘色長條（5000 筆）都遠高於其對應的藍色長條（2500 筆）。這直觀地證明了數據規模對於 logit-based 指標的有效性至關重要。更多的數據顯著提升了各指標尋找穩定分類門檻的能力，從而大幅提高了發音錯誤檢測的效能。
2. **指標排名的「反轉效應」**: 圖 1 最核心的洞見在於它清晰地揭示了指標排名的「反轉」。在藍色長條中，kurtosis 和 skewness 是相對的領先者；然而在橘色長條中，它們的表現被 mean_logit_margin, evt_k3 等指標遠遠超越。這個「反轉」現象極具建設性，它告訴我們：

- **在數據量不足、信號可能充滿雜訊時**，logit 分佈的「形狀」(如峰度和偏度)是一個比絕對數值更穩健、更可靠的判斷依據。
- **當數據量充足、模型判斷更穩定時**，logit 的「區分度」(mean_logit_margin)和「峰值強度」(evt_k3)這些更直接的指標，則成為了最強大的分類特徵。

3. **頂尖指標的收斂**：在 5000 筆數據的結果中(橘色長條)，我們可以看到排名前四的指標 mean_logit_margin, evt_k3, kl_to_onehot, entropy_mean) 形成了性能非常接近的「第一梯隊」。這建設性地指出，儘管這些指標的計算方式各不相同(分別代表區分度、峰值信心、與理想分佈的差距、模型混亂度)，但它們都從不同側面有效地捕捉了「模型判斷的確定性」這一核心概念。這意味著未來的研究可以嘗試將這些頂級指標進行融合，以期達到更佳的性能。

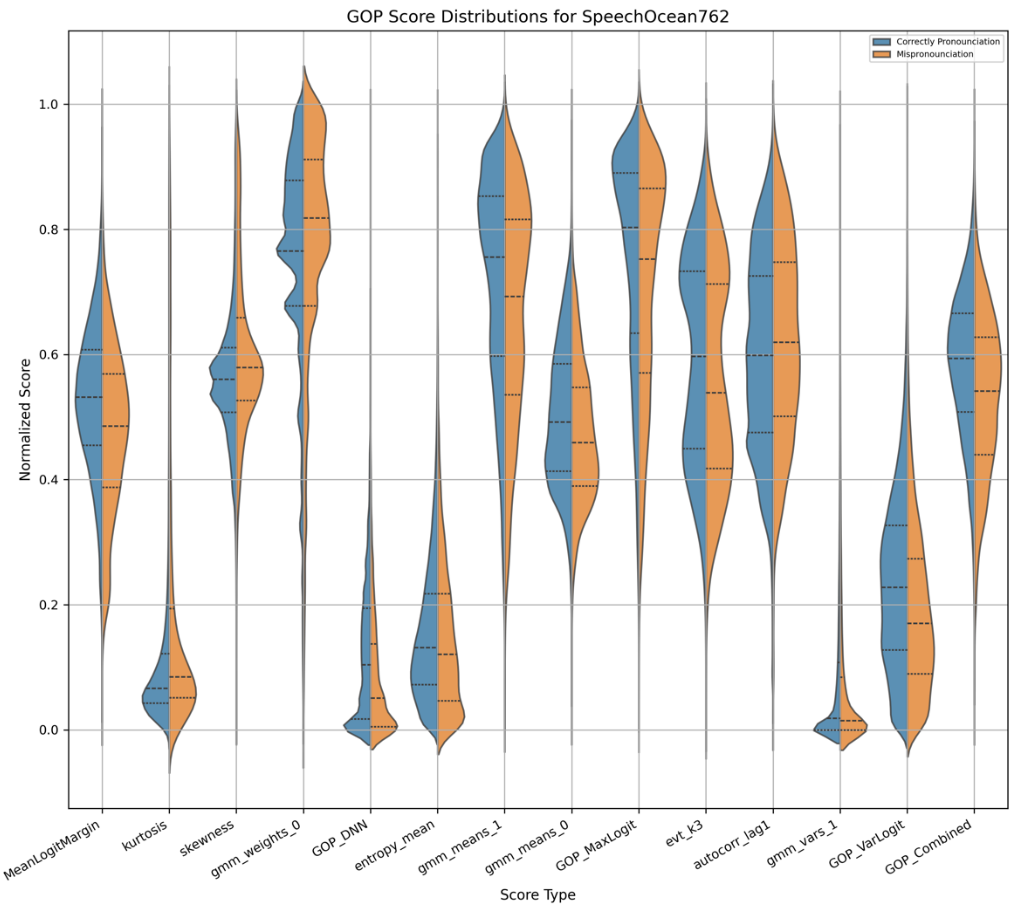


圖 1：Speechocean762 資料集 GOP 分數分佈的比較

4. 結論 (Conclusion)

本研究旨在系統性地擴展 logit-based GOP 分數的計算維度，探討超越傳統基本統計量(如最大值、變異數)的進階統計特徵，在發音錯誤檢測任務上的有效性。我們提出並實作了一套涵蓋分佈形狀(高階動差)、模型不確

定性（資訊理論）、時間穩定性（自相關分析）等多面向的指標，並透過在 SpeechOcean762 資料集的訓練集（2500 筆）與完整資料集（5000 筆）上進行的兩階段實驗，嚴謹地評估了這些指標的性能與穩定性。

我們的研究得出了一個核心且具指導性的結論：**logit-based GOP 指標的有效性與最佳選擇，高度依賴於實驗數據的規模。**

- 在數據量較少的初步實驗中，描述 logit 分佈「形狀」**的指標，特別是**峰度 (kurtosis)，展現出相對最佳的分類潛力。這表明在數據稀疏、雜訊較多的情況下，logit 分佈的異常形狀可能是比其數值大小更穩健的錯誤信號。
- 然而，在數據量加倍的完整實驗中，描述 logit **「區分度」**(mean_logit_margin) 和「峰值信心」 (evt_k3)的指標則逆轉成為表現最強的特徵。這證實了當有足夠的數據支撐時，模型對正確音素明確、高置信度的判斷，是區分發音正確與否最直接且有效的依據。

研究限制與未來展望 (Limitations and Future Work)

本次研究的主要限制是，我們僅評估了每種統計指標作為單一分類器的性能。基於本次的發現，我們規畫了幾個未來可能的研究方向：

1. **特徵融合建模**：最關鍵的下一步是將在完整資料集上表現最好的多個特徵（如 mean_login_margin, evt_k3, kl_to_onehot 等）**融合**起來，共同作為一個更強大的機器學習分類器（例如邏輯迴歸、梯度提升樹等）的輸入。我們預期這種多維度的綜合判斷，其性能將顯著超越任何單一特徵。
2. **數據規模的深入探討**：未來的研究可以進一步探討「形狀」指標與「區分度」指標發生性能交叉的數據量級，這有助於為不同規模的發音評估任務，提供自適應的特徵選擇策略。
3. **跨模型與跨語言的泛化性驗證**：本研究的發現基於單一聲學模型，未來可將這些統計指標應用於不同的模型架構（如 Whisper）或不同母語背景的學習者，以驗證我們結論的泛化能力。

總而言之，本研究不僅系統性地評估了一系列創新的 logit-based GOP 指標，更重要的是揭示了數據規模對指標選擇的關鍵影響，並為未來的發音評估研究，從「尋找單一最佳指標」轉向「根據條件進行多指標融合」，提供了堅實的實證基礎。

References

- Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, Helmer Strik. *Evaluating Logit-Based GOP Scores for Mispronunciation Detection, volume 1*. Interspeech 2025.
- Bi-Cheng Yan, Jiun-Ting Li, Yi-Cheng Wang, Hsin-Wei Wang, Tien-Hong Lo, Yung-Chang Hsu, Wei-Cheng Chao, Berlin Chen, *An effective pronunciation assessment approach leveraging hierarchical Transformers and pre-training strategies*, the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Bangkok, Thailand, August 11-16, 2024. (Long Paper)
- Tzu-Hsuan Yang, Yue-Yang He, Berlin Chen, *JCAPT: A Joint Modeling Approach for CAPT*, ISCA SLaTE-2025 Workshop.
- Yassine El Kheir, Ahmed Ali and Shammur Absar Chowdhury, *Automatic Pronunciation Assessment -- A Review*, EMNLP Findings(2023)
- Sandra Kanters, Catia Cucchiarini, Helmer Strik, *The Goodness of Pronunciation Algorithm: a Detailed Performance Study*, ISCA SLaTE-2

Learning Common Interests for Cold-Start Group Recommendation

Yu-Ting Cheng
National Taiwan Normal University, Taiwan
kirstin.yt.cheng@gmail.com

Pin-Hsin Hsiao
National Taiwan University, Taiwan
henry1233815@gmail.com

Chiou-Shann Fuh
National Taiwan University, Taiwan
fuh@csie.ntu.edu.tw

Pu-Jen Cheng
National Taiwan University, Taiwan
pjcheng@csie.ntu.edu.tw

Abstract

Previous studies on recommender systems have primarily focused on learning implicit preferences from individual user behaviors or enhancing recommendation performance by identifying similar users. However, in real-life scenarios, group decision-making is often required, such as when a group of friends decides which movie to watch together. Thus, discovering common interests has become a key research issue in group recommendation.

The most straightforward approach to group recommendation is to model the past joint behaviors of a user group. Nevertheless, this method fails to handle newly formed groups with no historical interactions. To address this limitation, we apply Graph Convolution Networks to capture high-order structural features within the user–item interaction graph, thereby uncovering the potential common interests of cold-start groups. Experimental evaluations on three real-world datasets demonstrate the feasibility and effectiveness of the proposed method.

Keywords: Discovery of Common Interests, Cold-Start Groups, Group Recommendation

1 Introduction

Recommender systems have become an essential component of modern digital experiences, assisting users in exploring products and potential social connections by analyzing their behaviors and preferences. For example, platforms such as Amazon and TripAdvisor provide personalized product and hotel suggestions based on user interactions and reviews.

Despite the impressive success of existing recommender systems in delivering personalized recommendations, they often overlook group decision-making scenarios, such as a group of friends choosing a movie to watch together or deciding on a restaurant for dining. Our work aims to bridge this gap by uncovering common interests within user groups, particularly for *cold-start groups* (a set of users who come together for the first time and for whom the system has no prior record of collective interactions or shared history). This capability not only enables recommendations that align with collective group preferences but also opens new possibilities for collaborative content creation, such as co-writing a script.

Prior research on group recommendation (Berkovsky, 2010; Baltrunas, 2010; Amer-Yahia, 2009) has primarily targeted *persistent groups*, in which members are fixed and have interacted multiple times as a group. In contrast, *cold-start group recommendation* poses a greater challenge, since *ephemeral groups* typically lack prior interactions or shared histories. In such cases, balancing individual preferences with group dynamics to produce recommendations that satisfy all members is a highly non-trivial problem.

Recent group recommendation advances (Sajjadi Ghaemmaghami, 2021) include attention-based aggregation over persistent groups (AGREE) (Cao, 2018) and multi-view modeling for occasional groups (GAME) (He, 2020a). For *ephemeral* groups without joint history, GroupIM (Sankar, 2020) maximizes

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 334–341
November 20-22, 2025

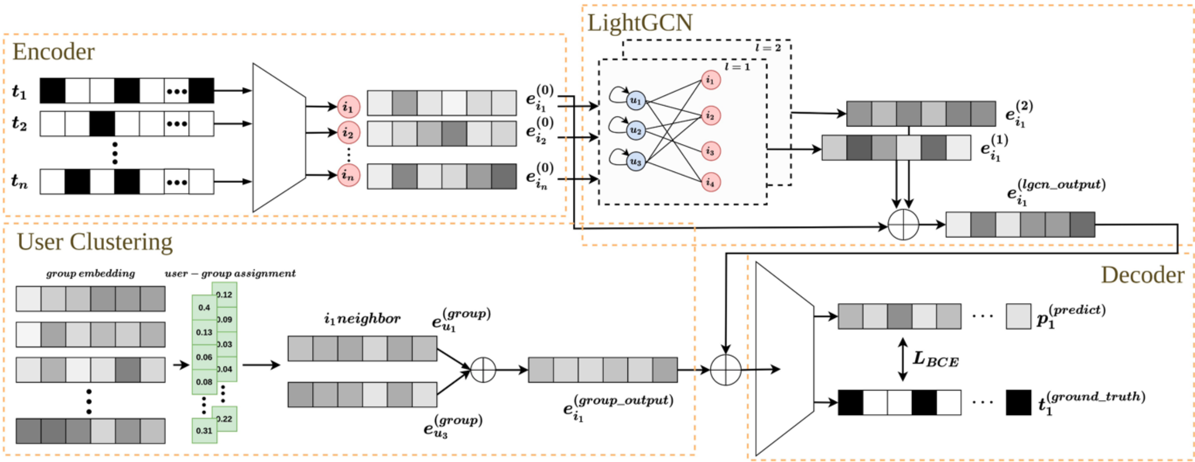


Figure 1: The COIN architecture illustrating training on item i_1 for user group $\{u_1, u_3\}$.

mutual information between group/user/item representations. These works generally lack interpretability and do not uncover the latent semantics behind group members' shared interests. Disentangled recommendation methods (Ma, 2019) are capable of learning factorized representations to capture latent semantics in user–item interaction data, but their focus remains on individual user behavior rather than on cold-start groups.

In this paper, we propose the **COmmon INtereset model (COIN)** to discover potential common interests in cold-start user groups. We leverage Graph Convolution Networks (GCNs) to capture high-order relations in user–item interactions and construct a *virtual item* for a cold-start group. This virtual item represents the most suitable recommendation for a given group, and by incorporating its tag attributes as auxiliary data (Liu, 2020), COIN can effectively reflect the group's potential common interests. The COIN model consists of four main components: (1) a **tag encoder** that transforms sparse item tag attributes into dense vectors; (2) **LightGCN** (He, 2020b), which captures high-order user–item interactions; (3) a **user clustering module** that models user–group–tag level preferences; and (4) a **tag decoder** that reconstructs tag-level semantics from the learned dense representations. Extensive experiments have been conducted on three real-world datasets, and the results demonstrate the feasibility and effectiveness of the proposed method.

2 The Proposed COIN Model

As shown in Fig. 1, the COIN model consists of four main components working together to uncover group-level common interests. First, the *Encoder* transforms sparse item tag attributes into dense embeddings, providing compact semantic representations for items. Next, *LightGCN* captures high-order relations in the user–item interaction graph, refining embeddings through graph propagation. Meanwhile, the *User Clustering* component models user–group–tag preferences by softly assigning users to latent groups and generating group-aware item embeddings. Finally, the Decoder combines the outputs from LightGCN and clustering, reconstructs tag semantics, and predicts the common interests of cold-start groups, optimized via a binary cross-entropy loss.

2.1 Problem Formulation

Let U denote the set of users, I the set of items, and T the set of tag attributes. Each item i is associated with a multi-hot vector t_i representing its tag attributes. We define the set of user–item interactions as $R^+ = \{(u, i) \mid u \in U, i \in I\}$. Given a *cold-start user group* S (a subset of users), the objective is to learn a function that predicts the top-k tag attributes representing the common interests of this cold-start group.

2.2 Tag Attribute Encoder

Each item tag attribute $t_i \in R^T$ is represented as a high-dimensional multi-hot vector. Since directly training with such sparse vectors is impractical, we employ a two-layer Multi-Layer Perceptron (MLP) as the encoder. The encoder projects each sparse tag vector into a dense embedding space, yielding an initial item embedding $e_i^{(0)} \in R^d$ for subsequent model training.

For users, we construct an embedding look-up table, where each column represents a user embedding $e_u^{(0)} \in R^d$. These user embeddings, together with the encoded item embeddings, serve as the foundation for later components of the COIN model.

The encoding process can be expressed as:

$$e_i^{(0)} = W_2 \cdot \mathrm{ReLU}(W_1 t_i),$$

$$E_u = [e_{u_1}^{(0)}, \ldots, e_{u_N}^{(0)}],$$

where W_1 and W_2 are trainable weight matrices, and E_u denotes the collection of user embeddings.

2.3 LightGCN

After obtaining the initial embeddings of users and items from the encoder, the next step is to capture their potential common interests. When multiple users interact with the same item, it indicates they may share latent preferences reflected in the item's tag attributes.

We leverage graph convolution network-based solutions, particularly LightGCN, which has proven highly effective for various recommendation tasks. The encoder's initial embeddings serve as the input. LightGCN propagates information across the user–item interaction graph:

$$e_u^{(l+1)} = \sum_{i \in N_u} \frac{1}{\sqrt{|N_u|}\sqrt{|N_i|}} e_i^{(l)},$$

$$e_i^{(l+1)} = \sum_{u \in N_i} \frac{1}{\sqrt{|N_u|}\sqrt{|N_i|}} e_u^{(l)},$$

where N_u denotes the set of items interacted with by user u, and N_i denotes the set of users interacting with item i.

However, high-order propagation may cause the problem of over-smoothing, where user embeddings lose their unique semantics and become dominated by item embeddings. To mitigate this issue, we introduce a residual connection for users, which preserves user-specific information:

$$e_u^{(l+1)} = e_u^{(l)} + \sum_{i \in N_u} \frac{1}{\sqrt{|N_u|}\sqrt{|N_i|}} e_i^{(l)},$$

$$e_i^{(l+1)} = \sum_{u \in N_i} \frac{1}{\sqrt{|N_u|}\sqrt{|N_i|}} e_u^{(l)}.$$

The discussion of the training–inference gap for items is deferred to Sections 2.6 and 2.7.

For efficiency in implementation and training, we rewrite the propagation rule in matrix form. Let $R \in R^{N \times M}$ denote the user–item interaction matrix, where $R_{ui}=1$ if user u has interacted with item i, and 0 otherwise. By adding user self-loops, the adjacency matrix is defined as:

$$A = \begin{pmatrix} I & R \\ R^\top & 0 \end{pmatrix},$$

with degree matrix D. Let $E^{(l)} \in R^{(N+M) \times d}$ be the embeddings at layer l. The propagation becomes:

$$E^{(l+1)} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} E^{(l)}.$$

After L layers of propagation, we aggregate embeddings from all of the layers (including the encoder's initial embeddings) by averaging, to retain semantic information learned at each stage:

$$e_i^{(\mathrm{lgcn\_output})} = \frac{1}{L+1} \sum_{l=0}^{L} e_i^{(l)}.$$

This ensures that the final item embeddings preserve both high-order relational knowledge and the semantic features from earlier layers.

2.4 User Clustering

We assume the existence of $|G|$ latent user groups, each capturing abstract and complex group–tag preferences. A user may belong to multiple groups simultaneously, and thus can be represented as a combination of these group memberships. Since no external resources are available, we adopt a simple soft clustering approach that is directly learned from user embeddings.

Formally, let $S \in R^{N\times|G|}$ denote the user–group assignment matrix, where $S_{i,j}$ corresponds to the probability of user u_i belonging to group g_j. To obtain this, we apply a linear projection $W_{proj} \in R^{d\times|G|}$ to the user embeddings, followed by a softmax function to ensure each row forms a valid probability distribution:

$$S = \text{softmax}(E_u W_{\text{proj}}).$$

Next, we maintain a group embedding look-up table, where each column represents a latent embedding $e_g \in R^d$. Rather than learning direct group–tag preferences (which would be computationally prohibitive given the high dimensionality of tag attributes), we instead learn a latent vector for each group. Combining the user–group assignment matrix with the group embeddings, we derive user representations in the group space:

$$E_g = [e_{g_1}, \ldots, e_{g_{|G|}}],$$
$$E_{\text{user\_group}} = SE_g.$$

Finally, an item embedding is represented by averaging over the group-based user embeddings of its neighboring users:

$$e_i^{(\text{group\_output})} = \frac{1}{|N_i|} \sum_{u \in N_i} e_u^{(\text{user\_group})}.$$

Through this design, the model captures group-level user preferences and mitigates the over-smoothing issue encountered in LightGCN for low-degree users.

2.5 Tag Attribute Decoder

To generate the final item representation, we apply a linear combination of the outputs from LightGCN and the user clustering module. Specifically, given hyperparameter α, the final item embedding is computed as:

$$e_i^{(\text{final})} = \alpha \cdot e_i^{(\text{lgcn\_output})} + (1-\alpha) \cdot e_i^{(\text{group\_output})}.$$

This embedding is then passed through a decoder, which mirrors the structure of the encoder. The decoder transforms the dense item embedding back into the tag attribute space, thereby reconstructing semantic features. Finally, we apply a *sigmoid* activation to produce probabilities for each tag attribute:

$$p_i = \sigma\big(W_4 \cdot \text{ReLU}(W_3 e_i^{(\text{final})})\big),$$

where W_3 and W_4 are trainable weight matrices. These probabilities are compared against the ground-truth tag attributes using binary cross-entropy loss, ensuring that the learned item embeddings preserve interpretable semantics aligned with observed tag data.

2.6 Model Training

During training, the COIN model jointly learns (1) user embeddings, (2) the encoder for projecting item tag attributes into a latent semantic space, (3) group embeddings with user–group assignments to capture group-level preferences, and (4) the decoder for reconstructing item tag attributes to reveal common interests.

To simulate cold-start groups, we adopt a masking strategy in which certain items are randomly removed, along with all co-interacted items from the same subset of users, thereby forming a common interest set. For instance, if users u_1 and u_3 interacted with items i_4 and i_7, both items are removed to define their shared interest during testing.

The model then predicts tag attribute distributions for items, which are optimized using binary cross-entropy (BCE) loss:

$$\mathcal{L}_{BCE} = \sum_{k=0}^{|T|} -\Big(t_{ik}\log(p_{ik}) + (1-t_{ik})\log(1-p_{ik})\Big),$$

Table 1: Statistics of the datasets

	# User	# Item	# Density	# Tag Number	# Avg Tag per item	# Group Avg User
citeulike-a	5551	16,980	0.00217	46,390	14.09	5.43
citeulike-t	7947	25,975	0.00065	52,946	11.19	2.04
yelp	15844	19042	0.00140	889	5.01	2.85

where t_{ik} is the ground-truth value of the k-th tag attribute for item i, and p_{ik} is the predicted probability. This ensures that reconstructed attributes closely align with true tag labels, guaranteeing that the learned embeddings preserve sufficient semantic information for decoding back into the tag attribute space.

2.7 Model Inference

During inference, the goal is to predict the probability distribution over tag attributes that represent the common interests of a given cold-start user group. As in training, the encoder first transforms item tag attributes into embeddings, producing both item and user representations.

To handle cold-start groups, we introduce a *virtual item* into the user–item interaction graph. This virtual item is connected to all users in the input group, enabling propagation through both LightGCN and the user clustering module. Unlike training, where the graph structure remains fixed, the inference graph is dynamically constructed for each cold-start group. The virtual item's embedding thus encodes the aggregated preferences of the group.

Finally, the virtual item embedding is decoded back into the tag attribute space, yielding probabilities for each tag. The *top-k tag attributes* are selected as the predicted common interests of the cold-start user group.

3 Experiments

In this section, we compare the proposed COIN model against several baselines and provide an empirical analysis of each model component.

3.1 Experimental Settings

We evaluate the COIN model on three publicly available real-world datasets: Citeulike-a, Citeulike-t (Wang, 2013), and Yelp, with their statistics summarized in Table 1. The Citeulike-a and Citeulike-t datasets are collected from *CiteULike*, an online platform where users share and manage academic papers. Each paper includes metadata such as title, abstract, and tag attributes, and in our experiments we utilize the user–paper interactions along with paper tag attributes to predict group-level common interests. The Yelp dataset is a subset of Yelp's business data, containing user–business interactions, reviews, check-ins, and location tag attributes. For our task, we specifically use the user–location interactions and location tag attributes to infer common interests within user groups.

Following the training and inference procedures described in Sections 2.6 and 2.7, we randomly mask certain items for testing. Items co-interacted by the same user subset are grouped together, and their averaged tag attributes are treated as the ground-truth common interests. For evaluation, we adopt three top-k ranking metrics (Krichene, 2020): Recall@20, F1@20, and NDCG@20. The predicted tag attributes of the virtual item are ranked and compared against the ground-truth top-k tag attributes.

We compare the COIN model against several baselines. The **Intersection** method is a naïve approach that defines a group's common interest as the intersection of all tag attributes associated with items interacted by group members. The **Probability** method adopts a simple probabilistic strategy, where each user's tag preference is calculated by multiplying the probability of the user interacting with an item and the tag probability of that item, and the group's common interest is then derived by combining the tag preferences of all group members. **MAGREE** is a modified version of the AGREE model (Cao, 2018), in which a

Table 2: Performance comparison

	citeulike-a			citeulike-t			yelp		
	Recall	F1	NDCG	Recall	F1	NDCG	Recall	F1	NDCG
Intersection	0.0002	0.0004	0.0019	0.0001	0.0003	0.0048	0.0015	0.0014	0.0013
Probability	0.0724	0.0987	0.2030	0.0336	0.0550	0.1825	0.2825	0.2436	0.4022
UAP	0.0570	0.0897	0.0234	0.0322	0.0600	0.3510	0.2597	0.2262	0.3640
MAGREE	0.0628	0.0978	0.2526	0.0355	0.0646	**0.3822**	0.2911	0.2454	0.3813
AE+GCN	0.0125	0.0184	0.0401	0.0018	0.0033	0.0166	0.0379	0.0314	0.0283
AE+LightGCN	0.0971	0.1379	0.2896	0.0617	0.1028	0.3431	0.2990	0.2477	0.4127
COIN	**0.1097**	**0.1517**	**0.3104**	**0.0646**	**0.1077**	0.3644	**0.3540**	**0.2930**	**0.4735**
Improvement	12.97 %	10.00 %	7.18 %	4.70 %	4.76 %	-4.88 %	8.36%	6.33%	6.32%

UAP: User Attribute Prediction Model MAGREE: Modified AGREE

tag attribute prediction layer is added to better suit our task. The **User Attribute Prediction Model (UAP)** trains only user embeddings and the decoder; item embeddings are generated during training by averaging the embeddings of interacting users and decoded to reconstruct tag attributes, while during inference a virtual item embedding is formed by averaging the embeddings of the input user set and then decoded to predict tag attributes. **AE+LightGCN** employs an encoder for item embeddings together with a user embedding lookup table, and applies LightGCN with residual connections to capture high-order semantics in the user–item graph, with the resulting embeddings decoded into tag attributes. Finally, **AE+GCN** (Kipf, 2017) serves as a variant using graph convolution networks, where each layer includes a linear transformation and activation function after graph propagation; this distinguishes it from LightGCN but also introduces potential mismatches between training and inference due to structural differences in the graphs.

3.2 Results and Discussion

The results are shown in Table 2, where "Improvement" indicates the relative gain of COIN over the second-best one (underlined).

The **Intersection** method, though intuitive, performs poorly since it only considers tag overlaps from user–item interactions, ignoring latent semantics and failing to generalize. The **User Attribute Prediction Model (UAP)** underperforms as well, since it does not learn item embeddings directly from tag attributes and relies on averaging user embeddings, which limits its ability to capture high-order semantics. **MAGREE** performs better than Intersection and UAP due to its attention mechanism, but still falls short of COIN because it does not explicitly model high-order relations in the user–item graph.

AE+GCN performs even worse than UAP: while it leverages an encoder and high-order interactions, its reliance on transformations and activations creates mismatches between training and inference, leading to degraded performance. **AE+LightGCN** achieves stronger results by learning item embeddings through an encoder and exploiting high-order relations without transformation mismatches; however, its performance drops on low-degree interactions where embeddings are prone to over-smoothing. By contrast, the **COIN** model consistently outperforms all baselines. Through tag-aware encoding, residual-enhanced LightGCN propagation, soft clustering to address over-smoothing, and effective decoding into tag attributes, COIN captures both high-order relations and latent semantics, leading to more accurate identification of common interests within groups.

3.3 Ablation Study

To evaluate the contribution of each component in the COIN model, we conduct an ablation study on two key modules: the residual connection in LightGCN and user clustering. The results are shown in Table 3.

Table 3: Ablation study

	citeulike-a			citeulike-t			yelp		
	Recall	F1	NDCG	Recall	F1	NDCG	Recall	F1	NDCG
COIN	**0.1097**	**0.1517**	**0.3104**	**0.0646**	**0.1077**	**0.3644**	**0.3546**	**0.2900**	**0.4735**
w/o residual	0.0876	0.1263	0.2677	0.0585	0.0986	0.3495	0.3446	0.2881	0.4698
w/o clustering	0.1009	0.1428	0.2942	0.0635	0.1055	0.3498	0.3179	0.2588	0.4318

Table 4: Impact of clustering over low-degree users

	citeulike-a			citeulike-t			yelp		
	Recall	F1	NDCG	Recall	F1	NDCG	Recall	F1	NDCG
w/o clustering	0.0944	0.1017	0.2380	0.0709	0.0865	0.2724	0.3104	0.2107	0.4013
with clustering	0.1108	0.1145	0.2698	0.0751	0.0901	0.2829	0.3674	0.2530	0.4643
Improvement	17.37%	13.36%	12.58%	5.92%	4.16%	3.85%	18.36%	20.07%	15.69%

Table 5: Impact of clustering over high-degree users

	citeulike-a			citeulike-t			yelp		
	Recall	F1	NDCG	Recall	F1	NDCG	Recall	F1	NDCG
w/o clustering	0.0991	0.1214	0.3553	0.0366	0.0573	0.5322	0.3027	0.2357	0.4469
with clustering	0.1061	0.1304	0.3775	0.0372	0.0579	0.5606	0.3333	0.2613	0.4878
Improvement	7.06%	7.04%	6.27%	1.63%	1.04%	5.33%	10.10%	10.86%	9.15%

Table 6: Impact of layer number

	citeulike-a			citeulike-t			yelp		
	Recall	F1	NDCG	Recall	F1	NDCG	Recall	F1	NDCG
COIN-1	0.0492	0.0782	0.2156	0.0494	0.0048	0.0508	0.0507	0.0521	0.0814
COIN-2	**0.1097**	**0.1517**	**0.3104**	0.0646	0.1077	**0.3644**	**0.3540**	**0.2930**	**0.4735**
COIN-3	0.0984	0.1382	0.2777	0.0636	0.1064	0.3643	0.3240	0.2668	0.4396
COIN-4	0.0997	0.1399	0.2828	**0.0665**	**0.1082**	0.3545	0.3078	0.2634	0.4388

The residual connection improves performance by directly preserving user-specific information during propagation. Without it, user embeddings tend to collapse into overly similar representations dominated by neighboring items, leading to a loss of individual semantics. As illustrated in Table 3, removing the residual connection reduces Recall and NDCG across all datasets, confirming that retaining user individuality is crucial for accurate recommendation.

The user clustering module also plays an important role. By softly assigning users to multiple latent groups, clustering captures higher-level group semantics and alleviates sparsity in low-degree interactions. This effect is evident in Table 3, where removing clustering leads to a noticeable performance drop, particularly for datasets with many low-degree users (e.g., CiteULike-t). Clustering enables the model to infer preferences for sparse users by leveraging patterns shared with similar users, thereby mitigating cold-start and over-smoothing issues.

We also investigate the impact of user clustering on groups with different interaction degrees. To this end, users are divided into low-degree and high-degree categories based on whether their number of interactions falls below or above the median interaction count in the dataset. The results, presented in Tables 4 and 5, show that user clustering improves performance in both categories but provides a more substantial gain for low-degree users. This demonstrates that clustering is particularly effective in alleviating the over-smoothing problem in sparse interaction scenarios, as it allows low-degree users to leverage shared group-level semantics to compensate for limited individual interactions.

3.4 Impact of Layer Number

COIN benefits from light propagation, which enables it to capture high-order interactions within the user–item graph. To evaluate the effect of different propagation depths, we tested configurations with 1, 2, 3, and 4 layers, and the results are summarized in Table 6. The findings show that COIN-2 and COIN-3 achieve the strongest performance across most metrics. In contrast, COIN-1, which does not include propagation, performs significantly worse, indicating that modeling high-order interactions is essential for learning meaningful common interests. However, when the number of propagation layers becomes too large, as in COIN-4, the model suffers from over-smoothing, where embeddings lose their distinctiveness and fail to capture nuanced user–item semantics. This suggests that a moderate depth, specifically two or three layers, strikes the best balance between leveraging high-order relationships and preserving semantic richness in the embeddings.

4 Conclusions

In this paper, we propose the COIN model for discovering common interests in cold-start user groups. Unlike prior methods focused on persistent groups, COIN tackles the problem of cold-start group recommendation by leveraging item tag attributes and high-order semantics captured through LightGCN and user clustering. Experiments on Citeulike-a, Citeulike-t, and Yelp datasets demonstrate that COIN consistently outperforms baseline methods. Ablation and sensitivity analyses confirm that residual connections reduce over-smoothing, clustering enhances low-degree user modeling, and decoding provides interpretable tag-level explanations.

Future work could extend the model by incorporating temporal dynamics to capture evolving group preferences and integrating multimodal signals such as reviews or images for richer semantics.

References

Amer-Yahia, S., Roy, S. B., Chawlat, A., Das, G., & Yu, C. 2009. Group recommendation: Semantics and efficiency. In *Proc. of the VLDB Endowment*.

Baltrunas, L., Makcinskas, T., & Ricci, F. 2010. Group recommendations with rank aggregation and collaborative filtering. In *Proc. of ACM Conference on Recommender Systems* (pp. 119-126).

Berkovsky, S., & Freyne, J. 2010. Group-based recipe recommendations: analysis of data aggregation strategies. In *Proc. of ACM Conference on Recommender Systems* (pp. 111-118).

Cao, D., He, X., Miao, L., An, Y., Yang, C., & Hong, R. 2018. Attentive group recommendation. In *Proc. of ACM SIGIR Conference* (pp. 645-654).

He, Z., Chow, C. Y., & Zhang, J. D. 2020a. GAME: Learning graphical and attentive multi-view embeddings for occasional group recommendation. In *Proc. of ACM SIGIR Conference* (pp. 649-658).

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. 2020b. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proc. of ACM SIGIR Conference* (pp. 639-648).

Kipf, T. N. 2017. Semi-supervised classification with graph convolutional networks. In *Proc. of* International Conference on Learning Representations.

Krichene, W., & Rendle, S. 2020. On sampled metrics for item recommendation. In *Proc. of ACM SIGKDD Conference* (pp. 1748-1757).

Liu, Q., Xie, R., Chen, L., Liu, S., Tu, K., Cui, P., ... & Lin, L. 2020. Graph neural network for tag ranking in tag-enhanced video recommendation. In *Proc. of ACM CIKM Conference* (pp. 2613-2620).

Ma, J., Zhou, C., Cui, P., Yang, H., & Zhu, W. 2019. Learning disentangled representations for recommendation. *Advances in Neural Information Processing Systems*, *32*.

Sajjadi Ghaemmaghami, S., & Salehi-Abari, A. 2021. DeepGroup: Group recommendation with implicit feedback. In *Proc. of ACM CIKM Conference* (pp. 3408-3412).

Sankar, A., Wu, Y., Wu, Y., Zhang, W., Yang, H., & Sundaram, H. 2020. GroupIM: A mutual information maximization framework for neural group recommendation. In *Proc. of ACM SIGIR Conference* (pp. 1279-1288).

Wang, H., Chen, B., & Li, W. J. 2013. Collaborative topic regression with social regularization for tag recommendation. In Proc. of *IJCAI* Conference (pp. 2719-2725).

Introduction: Persuasive Language in the Age of AI

Siaw-Fong Chung
Department of English
National Chengchi University
sfchung@nccu.edu.tw

Abstract

Persuasive language shapes communication across disciplines and everyday life. As large language models (LLMs) become increasingly integrated into these spheres, understanding persuasion now encompasses both human and machine discourse. This introduction examines how persuasive language operates across diverse contexts by analyzing the interactional frameworks of human and AI communication. It also explores how persuasion emerges in human-AI exchanges and how these insights can inform language education and communication practices. Drawing on perspectives from linguistics, computer science, journalism, and communication studies, it presents persuasion as both a rhetorical and interactional process shaped by technology. Ultimately, it aims to deepen understanding of how AI transforms persuasive practices and to promote greater awareness of persuasion in language learning.

Keywords: Persuasion, persuasive language, AI, AI-human interaction, discourse

1. The Language of Persuasion

Dillard & Pfau (2002:x), in their edited book *The Persuasion Handbook*, outlined the broad scope of persuasion. In this introduction, we present key definitions of persuasion and the linguistic resources that underpin it, followed by a discussion of persuasive attempts, which refer to the strategies and linguistic techniques used to influence others in ordinary discourse as well as in human-AI (and vice versa) interaction.

Regarding the basic conceptual concerns and definitions of persuasion, several dimensions can be identified. The language of persuasion can first be examined from the receiving end—the state of being persuaded:

> Thus, the phrase 'being persuaded' applied to situations where behavior has been modified by symbolic transactions (messages) that are sometimes, but not always, linked with the coercive force (indirectly coercive) and that appeal to the reason and emotions of the person(s) being persuaded. (Miller, 2002: 7)

The coercive force of language is often indirect. It gives persuasive messages power to make people change behavior, attitude, or belief without an explicit threat. For this reason, it is often linked to the use of persuasion strategies. These strategies refer to techniques employed to influence the persuadee's decision to align with the persuader's goal. When the process succeeds, the persuadee is said to be persuaded. Miller (2002) further explains that persuasion typically involves behavioral conversion, meaning the abandonment of one course of action and the adoption of another. In discussing the notion of being persuaded, Marwell and Schmitt (1967, as cited in Miller, 2002: 5) identified sixteen strategies, among which 'promise', 'threat', and 'aversive stimulation' have been said to "derive their effectiveness from the persuader's ability to dispense rewards or mete out punishments to the intended persuadee(s)." This means that these strategies aim to use rewards or punishments as ways to make the persuadee agree with the persuader. Other strategies requiring "social rewards resulting from compliance" are 'moral appeal', 'altruism' (i.e., willingness to do things that bring advantages to others, even if it results in disadvantage for yourself), 'esteem positive' (positive self), and 'esteem negative' (negative self). These require social approval, as do 'being respected,' 'being popular,' and 'being in' (cf. p. 5). Collectively, these strategies engage the persuadee's need for social acceptance and conformity to the persuader's intended action. In this sense, the coercive force of language operates not through overt control but through strategies that subtly manipulate social values and psychological needs, giving persuasion its enduring power.

Hosman (2002), in the same volume, emphasizes that one crucial element of persuasion is language itself. The examination of strategies and their correspondence to linguistic

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 342–347
November 20-22, 2025 ©2025 Association for Computational Linguistics

features can be observed in Shih et al. (2021) on the annotation of propaganda techniques in Chinese political news texts. The authors identified several persuasive techniques used in Chinese newspapers to achieve political purposes. English examples based on Da San Martino et al. (2019) were also provided in their paper.

However, identifying the strategies is not the only way to analyze the language of persuasion. Persuasion can also occur through the way concepts are defined and framed, since definitions themselves can shape attitudes and influence judgments. As early as 1944, Stevenson proposed the theory of persuasive definition— 'a definition can be effective as a device of deceptive persuasion' (cited in Walton, 2005: 162). In the following dialogue on culture analyzed by Stevenson (1944: 211), argumentation based on definition is shown:

(1) The Dialogue on Culture

A: *He has had but little formal education, as is plainly evident from his conversation. His sentences are often roughly cast, his historical and literary references rather obvious, and his thinking is wanting in that subtlety and sophistication which mark a trained intellect. He is definitely lacking in culture.*

B: *Much of what you say is true, but I should call him a man of culture notwithstanding.*

A: *Aren't the characteristics I mention the antithesis of culture, contrary to the very meaning of the term?*

B: *By no means. You are stressing the outward forms, simply the empty shell of culture. In the true and full sense of the term, ''culture'' means imaginative sensitivity and originality. These qualities he has; and so I say, and indeed with no little humility, that he is a man of far deeper culture than many of us who have had superior advantages in education.*

From this example, we observe only one aspect of how language can be used to persuade. We cited it because it serves as a classic illustration by Stevenson (1944), who was among the first to link persuasion theory to language use. Language, in general, encompasses tone, lexical choice, pragmatic strategies, and textual arrangement, all of which influence the effectiveness of persuasion. We will not cover every element in depth, but we will show how language shapes persuasion and communication with AI. Next, we explore corpus resources that help analyze persuasive language.

2. The Corpus of Persuasion

In this era, corpus collection has become increasingly common, and more shared linguistic resources are now available. In this introduction, we will survey existing corpora on persuasion that are accessible for use with appropriate acknowledgements and, where required, through consent or application. Corpora that are not available for use will not be included. Our initial step is to show the availability of existing English corpora.

Among the available corpus resources, a well-known series was developed by Walker and her colleagues at the Natural Language and Dialogue Systems Lab, University of California, Santa Cruz (https://nlds.soe.ucsc.edu/). These corpora contain naturally occurring dialogues rich in persuasive-strategy data, and are thus valuable for studying human-human and human-AI interaction. Walker et al. (2012a) established a corpus on deliberation and debate (see also Walker et al., 2012bc; Abbott et al., 2011), focusing on personality analysis and the styles of argumentation that resonate with different individuals. Although personality analysis is not the primary focus of our study on persuasion, the corpus provides dialogue-based persuasive language data that are valuable for linguistic analysis. In addition to the corpora developed by Walker and colleagues, other studies have created specialized datasets, though many remain unavailable to the public. The main accessible corpora are summarized in Table 1 of our survey. Although a large body of research exists on the automatic detection of persuasion, such studies fall beyond the scope of this section.

Table 1: List of Corpora on Debates or Persuasion

Corpora	Authors	Contents
The Persuasion and Personality Corpus	Lukin et al. (2017)	User-generated, factual vs. emotional dialogic exchanges compared to the effects on belief change to balanced, curated arguments.

The Internet Argument Corpus (IAC) version 2	Abbott et al. (2016); Walker et al. (2012a)	4forums (414K posts), ConvinceMe (65K posts), and a sample from CreateDebate (3K posts). It includes topic annotations, response characterizations (4forums), and stance
Persuasion For Good Corpus	Wang et al. (2019)	A collection of online conversations generated by Amazon Mechanical Turk workers, where one participant (the persuader) tries to convince the other (the persuadee) to donate to a charity. This dataset contains 1017 conversations, along with demographic data and responses to psychological surveys from users. 300 conversations also have per-sentence human annotations of dialogue acts that pertain to the persuasion setting, and sentiment.
The Multimodal Persuasive Dialogue Corpus	Kawano et al. (2022)	60 subjects (43 males and 17 females) between 18 and 38 years old for a dialogue experiment with the humanoid android ERICA
A Persuasive Dialogue Corpus	Hiraoka et al. (2014)	Dialogue between 3 professional salespeople and 19 subjects, where the salesperson is trying to convince a customer to buy a particular product.
ParlaMint corpora: 17 corpora of parliamentary debates	Erjavec et al. (2021)	A collection of 17 multilingual comparable corpora consisting of parliamentary debates. The ParlaMint corpora include debates of 17 national parliaments: Bulgarian parliament, Belgian parliament (French and Dutch language), British parliament (English language) Czech parliament, Croatian parliament, Danish parliament, Dutch parliament, French parliament, Hungarian parliament, Icelandic parliament, Italian parliament, Latvian parliament, Lithuanian parliament, Polish parliament, Slovenian parliament, and Spanish parliament.
VivesDebate	Ruiz-Dolz et al. (2021)	An Annotated Multilingual Corpus of Argumentation in a Debate Tournament
United Nations General Debate Corpus (UNGDC)		Texts of General Debate statements from 1970 (Session 25) to 2016 (Session 71)

Corpus collection has become increasingly common, accompanied by the growing availability of shared resources. In Taiwan, there are also several persuasion-related corpora, though most remain private. Research combining persuasion and AI remains largely unexplored. Next, we will discuss whether AI has an underlying philosophy.

3. The Underlying Philosophy of AI

What is the underlying philosophy of AI? Does it have one? The answer, perhaps surprisingly, is yes. AI embodies traces of human thought because it is built upon a vast collection of human-written materials. In the following section, we elaborate on this idea and consider how human perspectives and biases become embedded in AI systems.

We examine the underlying interaction mechanism framework of persuasion to explore how language is used in the process and how interactions between AI and humans can be applied to language teaching and other contexts. The term underlying interaction mechanism framework is used in a sense similar to what some scholars call schemas, frames, or scripts, defined as follows:

> Frames and scripts are constructs which were originally developed by researchers in the field of artificial intelligence. The constructs made it possible to represent in computer memory those aspects of world knowledge which appear to be involved in the natural processing of texts. […] According to de Beaugrande and Dressler (1981:90), frames constitute 'global patterns' of 'common sense knowledge about some central concept', such that the lexical item denoting the concept typically evokes the whole frame. In essence, frames are static configurations of knowledge. Scripts are associated with […] basic level events such as 'do the washing' and 'visit the doctor', which are structured according to the expected sequencing of expected events (cf. Rosch 1978). (Taylor, 1995:89, italics added)

From this excerpt, our notion of the underlying interaction mechanism includes both frames and scripts, representing the static and dynamic aspects mentioned by Taylor. Our focus, however, is on the mechanisms that shape interaction between humans and machines. The "common-sense knowledge about some central concept" (de Beaugrande and Dressler, 1981: 90) reflects configurations of world knowledge "structured according to the expected sequencing of events" (Rosch, 1978; Taylor, 1995: 89). This perspective is also reflected in our debate chatbot project, where interactional patterns emerge dynamically through turn-taking and topic development. It further relates to the study by Yen and Chung (2025, this volume), which showed how discourse markers function as cues for coherence, stance, and engagement in human-AI dialogue. Their goal is to use AI to help students practice the language of persuasion. During the chatbot activities, they collected and analyzed the AI's responses and added them to a new corpus section, the Corpus of Persuasion (Interaction with AI). This dataset will then be compared with the existing human-to-human corpus to examine whether AI demonstrates similar goals, reasoning patterns, or underlying philosophical tendencies.

Müller (2025) outlined the principal topics, arguments, and positions in the philosophy of AI, excluding ethical concerns. He argued that beyond intelligence and computation, it is essential to view AI through the lens of cognition, encompassing "perception, action, meaning, rational choice, free will, consciousness, and normativity." These dimensions provide a useful foundation for our investigation into the kinds of cognition AI demonstrates when engaging in persuasive language. Similarly, Hipólito (2023) emphasized that AI is deeply rooted in human sociocultural practices. Building on this view, our line of research calls for greater attention to the underlying philosophies that shape how AI generates persuasive responses. Since technologies are inherently designed with human values and practices, understanding what has been "taught" to AI requires direct interaction with it.

In this section, we have shown two key points. First, AI systems display preferences shaped by the materials on which they are trained; their underlying philosophy reflects human values and practices. Second, to understand AI's philosophy in its use of persuasive language, we must communicate with it through activities such as debates or argumentation.

4. Application of AI in the Classroom

Many studies have incorporated AI into classroom learning. Su et al. (2023) used ChatGPT to teach argumentative writing, examining prompt design and the changes students made before and after editing with AI assistance. Lin (2022) investigated how students' positions influence argumentation learning across online and face-to-face environments. Chalaguine and Hunter (2019) argued that chatbots should be trained to understand both sides of an argument, including conflicting viewpoints, so that they can handle controversial topics and formulate appropriate responses. This insight is particularly relevant for classroom practice. When integrating AI into lessons, it is crucial that chatbots be responsive to multiple perspectives. Only then can students learn effectively by engaging in meaningful, dynamic exchanges with the machine.

As these studies suggest, preparing both students and teachers for this new era is essential. Since AI carries an underlying philosophy shaped by human values and reasoning, understanding how to integrate it thoughtfully in educational contexts becomes even more critical. The use of AI in classrooms is no longer a reversible trend, and educators must learn not only its tools but also its underlying assumptions. As Gillani et al. (2023: 99) noted, "AI is a loose umbrella term that refers to a collection of methods, capabilities, and limitations." Building on this view, applications of AI in education can be seen in areas such as: (a) intelligent tutoring systems (ITS);
(b) assessment and feedback;
(c) coaching and counselling;
(d) school-choice suggestions; and
(e) outcome prediction.

Although these uses are powerful, Gillani et al. also cautioned that it is the values embedded in AI that affect the outcomes we receive:

> All technologies (including those powered by AI) have been designed with a set of values, practices, and use-cases in mind—and therefore, can be changed, even if they appear opaque or difficult to understand. (Gillani et al., 2023: 107)

This view that AI embodies values and a form of philosophy is important because it demonstrates how AI mirrors human practices. This aligns with our view that AI is not value-neutral but reflects an underlying philosophy shaped by human reasoning and social practice. Recognizing these human imprints is essential when integrating AI into education, as they determine not only how technologies function but also how teachers and learners interact with them. In other words, it may have a philosophy shaped by the materials used to build it. Boddington (2023a) in *Philosophy for AI Ethics*, and again in Boddington (2023b), emphasized that we must understand humans before we can understand AI ethics:

> Issues outlined of relevance to AI ethics include questions concerning the place of human beings in the natural world; claims of particular roles that humans may have; claims that human beings have some essential nature; claims about the relationship of humans to the mind and to embodiment; the boundaries and limits to human nature; and claims about divisions within human nature, our strengths and weaknesses, and how humans may be improved. (Boddington, 2023b, abstract)

These reflections bring us back to the central idea that understanding AI begins with understanding ourselves. If AI mirrors human reasoning and social practices, then its ethical and philosophical dimensions are extensions of our own. As Boddington recalled, any exploration of AI's nature must start from the study of what it means to be human, including our cognition, morality, and limitations. For educators and researchers alike, this means that integrating AI responsibly involves not only technical competence but also philosophical awareness. In recognizing that AI inherits the values embedded in human knowledge, we affirm the need for continual reflection on how our creations think, reason, and persuade on our behalf. In the era of AI-mediated persuasion, understanding the mechanics of how machines generate and deploy persuasive language becomes paramount for guiding informed educational practices.

Acknowledgements

This study was supported by the research project On Building the Language of Persuasion Corpus and Analyzing the Underlying Interaction Mechanism Frameworks: A Study of Generative AI and its Educational Applications (Project No. 113-2410-H-004-064-MY2).

References

Abbott, Rob, Brian Ecker, Pranav Anand, & Marilyn A. Walker. 2016. "Internet Argument Corpus 2.0: An SQL Schema for Dialogic Social Media and the Corpora to Go with It." In *Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Abbott, Rob, Marilyn A. Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowman, & Joseph King. 2011. "How Can You Say Such Things?!?: Recognizing Disagreement in Informal Political Argument." In *Proceedings of the Workshop on Language in Social Media (LSM)*, Portland, Oregon, USA.

Boddington, P. 2023a. Philosophy for AI Ethics: Metaethics, Metaphysics, and More. In *AI Ethics. Artificial Intelligence: Foundations, Theory, and Algorithms*. Springer, Singapore. https://doi.org/10.1007/978-981-19-9382-4_7

Boddington, P. 2023b. Humans and Intelligent Machines: Underlying Values. In *AI Ethics. Artificial Intelligence: Foundations, Theory, and Algorithms*. Springer, Singapore. https://doi.org/10.1007/978-981-19-9382-4_5

Chalaguine, L. A. & A. Hunter. 2019. "Knowledge Acquisition and Corpus for Argumentation-Based Chatbots." *The 3rd Workshop on Advances in Argumentation in Artificial Intelligence, co-located with the 18th International Conference of the Italian Association for Artificial Intelligence (AIIA)*, Volume 2528.

Da San Martino, G., S. Yu, A. Barrón-Cedeño, R. Petrov, & P. Nakov. 2019. "Fine-Grained Analysis of Propaganda in News Articles." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5636–5646. https://doi.org/10.18653/v1/D19-1565

De Beaugrande, R. & W. Dressler. 1981. *Introduction to Text Linguistics*. London: Longman.

Dillard, James Price & Michael Pfau. 2002. *The Persuasion Handbook: Developments in Theory and Practice*. Thousand Oaks, London, & New Delhi: Sage.

Erjavec, Tomaž et al. 2021. "Linguistically Annotated Multilingual Comparable Corpora of Parliamentary Debates ParlaMint.ana 2.1." *Slovenian Language Resource Repository CLARIN.SI*, ISSN 2820-4042. http://hdl.handle.net/11356/1431

Gillani, N., R. Eynon, C. Chiabaut, & K. Finkel. 2023. "Unpacking the 'Black Box' of AI in Education." *Educational Technology & Society*, 26(1), 99–111.

Hipólito, Inês. 2023. "The Human Roots of Artificial Intelligence." *PsyArXiv*, May 2. doi:10.31234/osf.io/cseqt.

Hiraoka, Takuya, Graham Neubig, Sakriani Sakti, Tomoki Toda, & Satoshi Nakamura. 2014.

"Construction and Analysis of a Persuasive Dialogue Corpus." *International Workshop on Spoken Dialogue Systems Technology*.

Hosman, Lawrence A. 2002. "Language and Persuasion." In Dillard, James Price & Michael Pfau (eds.), *The Persuasion Handbook: Developments in Theory and Practice* (pp. 371–389). Thousand Oaks, London, & New Delhi: Sage.

Lin, Y. R. 2022. "The Influence of Students' Position on Argumentation Learning Through Online and Face-to-Face Environments." *International Journal of Science Education*, 44(17), 2632–2657.

Lukin, Stephanie, Pranav Anand, Marilyn A. Walker, & Steve Whittaker. 2017. "Argument Strength Is in the Eye of the Beholder: Audience Effects in Persuasion." *15th European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, Spain.

Marwell, G. & D. R. Schmitt. 1967. "Dimensions of Compliance-Gaining Behavior: An Empirical Analysis." *Sociometry*, 30, 350–364.

Miller, Gerald R. 2002. "On Being Persuaded: Some Basic Distinctions." In Dillard, James Price & Michael Pfau (eds.), *The Persuasion Handbook: Developments in Theory and Practice* (pp. 3–16). Thousand Oaks, London, & New Delhi: Sage.

Misra, A. & M. A. Walker. 2013. "Topic-Independent Identification of Agreement and Disagreement in Social Media Dialogue." In *14th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL)*, Metz, France.

Müller, Vincent C. 2025. "Philosophy of AI: A Structured Overview." In Nathalie A. Smuha (ed.), *Cambridge Handbook on the Law, Ethics and Policy of Artificial Intelligence*. Cambridge: Cambridge University Press, 1–25.

Rosch, E. 1978. "Principles of Categorization." In E. Rosch & B. B. Lloyd (eds.), *Cognition and Categorization*, 27–48. Hillsdale: Lawrence Erlbaum.

Ruiz-Dolz, Ramon, Montserrat Nofre, Mariona Taulé, Stella Heras, & Ana García-Fornes. 2021. "VivesDebate: A New Annotated Multilingual Corpus of Argumentation in a Debate Tournament." *Applied Sciences*, 11(15): 7160. https://doi.org/10.3390/app11157160

Shih, Meng-Hsien, Ren-Feng Duann, & Siaw-Fong Chung. 2021. "The Analysis and Annotation of Propaganda Techniques in Chinese." *Computational Linguistics and Chinese Language Processing*, 26(1), 79–104.

Stevenson, C. L. 1944. *Ethics and Language*. New Haven: Yale University Press.

Su, Y., Y. Lin, & C. Lai. 2023. "Collaborating with ChatGPT in Argumentative Writing Classrooms." *Assessing Writing*, 57, 100752.

Taylor, John R. 1995. *Linguistic Categorization: Prototypes in Linguistic Theory*. 2nd ed. Oxford: Oxford University Press.

Walker, Marilyn A., Pranav Anand, Jean E. Fox Tree, Rob Abbott, & Joseph King. 2012a. "A Corpus for Research on Deliberation and Debate." In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Walker, M. A., P. Anand, R. Abbott, & R. Grant. 2012b. "Stance Classification Using Dialogic Properties of Persuasion." In *3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, Seogwipo, Korea.

Walker, M. A., P. Anand, R. Abbott, J. E. Fox Tree, C. Martell, & J. King. 2012c. "That's Your Evidence?: Classifying Stance in Online Political and Social Debate." *Decision Support Sciences*, 1–30.

Walton, D. 2005. "Deceptive Arguments Containing Persuasive Language and Persuasive Definitions." *Argumentation*, 19, 159–186. https://doi.org/10.1007/s10503-005-2312-y

Wang, Xuewei, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, & Zhou Yu. 2020. "Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good." *Proceedings of ACL*.

Yen, Yu-Che & Siaw-Fong Chung. 2025. "Stance and Cohesion: The Use of However and While in AI-Human Argumentative Discourse." *The 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*. National Taiwan University, Taipei.

Stance and Cohesion: The Use of *However and While* in AI-Human Argumentative Discourse

Yu-Che Yen
Department of English
National Chengchi University
yantwstudy@gmail.com

Siaw-Fong Chung
Department of English
National Chengchi University
sfchung@nccu.edu.tw

Abstract

This study investigates how connectives *However* and *While*, signaling contrast/ concession to construct stances, are distributed by AI chatbots in task-based argumentations. The corpus, comprising 13,482 words of chatbot-produced discourse, was analyzed to examine the connectives' sentence positions and their relation to content-, writer-, and reader-oriented propositions, based on an integrated framework of Hyland's (2005) framework and Thetela's (1997) evaluative-entity framework. A total of 124 tokens of *However* and *While* were extracted, excluding tokens whose stance and cohesive functions can't be clearly interpreted. Results show sentence-initial *However* (N=40) and sentence-initial *while* (N=59) are the primary devices for asserting a writer-oriented stance, signaling evaluation, claim or counter-claim. Sentence-initial *while* are more frequently used to frame a factual premise before projecting writer orientation. As to sentence-medial *while*, both preceding and subsequent clauses are often presented content-oriented propositions, indicating achieving cohesion is prioritized over expressing an evaluative stance. This study concludes that the use of these connectives, strategically applied in AI-human argumentations, shows how connectives contribute to manage stance construction and discourse coherence.

Keywords: Stance, Cohesion, AI, Argumentative, Persuasion

1. Introduction

Conjunctions, according to Fraser (1999), refer to its interchangeability as "discourse markers, discourse connectives, discourse operators, pragmatic connectives, sentence connectives, and cue phrases (p. 931), are used to establish cohesive relations within a context. Fraser asserts that these cohesive units are used to segment discourses as S1 and S2 to signal the relationship between propositions. Rehbein (2019) further emphasizes the importance of using specific form of connectives (concessive and contrastive ones) in either informal/ formal or spoken/ written discourse can present pragmatic strategies in making arguments more convincing and persuasive. Kamalski et al. (2006) found that in argumentative and persuasive discourse, explicit coherence marking provides not only clarities but also positive evaluations for interlocutors (also see Felder, 2015; Stab and Gurevych, 2014; Eckle-Kohler et al., 2015). Eckle-Kohler et al. (2015) found that discourse connectives are featured in signaling a prediction between premise and claim in argumentative context. The evidence from these empirical studies show that discourse connectives are used to make cohesion strategically. Also, Maamuujav (2025) found that rhetorical features of discourse connectives strongly help make predictions in argumentative context.

Among all these discourse connectives, the adversative relation plays a particular prominent role in sharing the argumentation. Halliday and Hasan (1976) assert that adversative connectives signal a proposition that is "contrary to expectation" (p. 252). Similar assertions can be found in Rehbein (2019). Rehbein found that discourse connectives (e.g., though, while, thus, however) often used in written articles. The discourse connective *however*, classified by Halliday and Hasan (1976) as an emphatic adversative form and described by Schiffrin (1987) as a contrastive marker, signals a semantic relationship of contrast that enables a writer or speaker to return to a previous concern or to preface a defense against an opposing view. By indicating contrast (*but*, *however*), addition (*moreover*), causation (*thus*), etc., writers manage the reader's comprehension along the argumentative path. A speaker or a writer can deploy pragmatic strategies in framing arguments with using these adversative/ contrastive connectives. Despite the established importance of these rhetorical devices, the systematic variation in how specific discourse relations, particularly adversative markers used to express argumentative stances, are deployed for persuasion strategies across different, controlled communication settings remains an area actively awaiting deeper empirical investigation. While

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 348–357
November 20-22, 2025

studies have established that the use of connective forms (e.g., *however* vs. *but*) is genre-specific, this study aims to explore whether variations at the level of the semantic discourse relation reflect different persuasive strategies.

2. Literature Review

2.1 Significance of Connectives for Coherence and Interaction

Halliday and Hasan's (1976) seminal work on cohesion classifies conjunction (connectives) as a key cohesive relation that specifies "the way in which what is to follow is systematically connected to what has gone before" (p. 227). In other words, connectives explicitly link propositions so that a text links together logically. Schiffrin (1987) likewise noted that discourse markers "have a role in accomplishing the integration needed for discourse coherence" by helping readers navigate transitions between ideas (p. 29). Beyond creating cohesion, connectives serve an interactive, reader-guiding function. They signal the intended relationship between statements and serve to cue readers how to interpret new information. For example, Fraser (1999) defined discourse markers as lexical expressions (often conjunctions or adverbs) that "signal a relationship between the interpretation of the segment they introduce, S2, and the prior segment, S1" (p. 950). For example, Fraser stated that discourse markers (e.g., *However, Furthermore, Thus, Incidentally*) mark segmentations either in canonical case or joined by a subordinate clause. Hyland's (2005) framework underscores this interactive role of connectives in writer reader engagement. The framework classifies logical connectives (also treated as transitions) as interactive resources that organize discourse for the reader's benefit. Such transitions "help readers interpret pragmatic connections between steps in an argument, signaling additive, causative, or contrastive relations in the writer's thinking" (Hyland, 2005, p. 59). In sum, connectives function as signposts that not only bind text together cohesively but also guide readers through the argument, which reflects the writer's effort to manage and even influence the audience. This guiding/ interactional function is especially crucial in persuasive contexts, where the effective use of connectives can clarify argumentative structure and subtly involve readers in the development of the discourse.

2.2 Contrastive Connectives and Stance in Persuasive Writing

Empirical research on persuasive discourse has emphasized the crucial role of contrastive connectives in structuring arguments and guiding reader interpretation. In a corpus study of persuasive texts, Rehbein (2019) demonstrated that even when author and topic are held constant, the frequency of specific contrastive markers varied remarkably with communicative context. For instance, connectives such as *however* and *while* were found to be far more typical of written, monologic articles, whereas a marker like *but* (along with causal connectives like *because* and *so*) dominated in dialogic interview settings. This distributional shift suggests that writers strategically adapt connective usage to fit audience expectations and medium constraints. Indeed, discourse connectives have been identified as strategic devices for persuasion: certain connectors reliably signal argumentative moves and even improve a text's reception. A recent large-scale study by Maamuujav (2025) on student essays found that higher-rated persuasive writing is characterized by adept use of rhetorical features such as logical connectives and cohesive structuring. The discourse connective, *thus*, for example, provides the cohesive structuring necessary to link the premise (S1) and the ensuing conclusion/research rationale (S2) into a single, interconnected semantic unit. It explicitly manages the argumentative path by cueing the reader that they are about to receive an inference based on the preceding information. In Maamuujav's study, a latent factor representing rhetorical cohesion significantly predicted holistic writing quality. These findings collectively shown that contrastive connectives (e.g. *however*, *although*, *while*, *but*) are not mere add-ons to argumentative texts but crucial to how writers articulate their stance and achieve persuasion.

Despite these empirical progresses, there are clear limitations of these studies (such as Maamuujav, 2025; Rehbein, 2019, Thetela, 1997) with regard to connective usage and writer stance. A notable gap is that many studies (such as Fraser, 1999; Rehbein, 2019) treat all contrastive connectives as functionally interchangeable signals of contrast or concession without examining how individual connectives might convey subtle evaluative stances. In broad quantitative analyses (such as Maamuujav, 2025), connectives are often tied together as a single cohesive device category. This implicitly assumes that choosing *however* versus *while* makes little difference so long as a contrast is signaled. Rehbein's (2019) study noted, discourse connectives are highly ambiguous and polyfunctional, being capable of expressing multiple types or flavors of relations. For instance, *however*, most commonly employed as a sentence-initial conjunctive adverb, typically signals a more forceful adversative stance by explicitly marking a shift to the writer's counter-argument. For example, in Fraser's (1999) study, the example "*Harry is old enough to drink. However, he can't because he has hepatitis.*" (p.938) showed that the connective *However,* acting as a cohesive tie connecting the message of S2 (the counter-argument) to S1 (the premise), introduces a counter-argument or correction, which is supported by its nature as a discourse marker and its structural independence. By contrast, *while*, when functioning as a subordinating conjunction in concessive use, often introduces a clause that mitigates the opposition and relegates the contrasting view to a backgrounded or less prominent position. For example, in Rehbein' (2019) study, the example "*Mary likes to read while John loves cooking.*" (p. 148) showed that the proposition violated the expectation, leading to an asymmetric connection. In other words, *however* overtly signals a rebuttal or correction, sharply delineating the writer's position against a preceding point, while a connective like *while* can mitigate opposition by weaving the contrast into the same sentence, thereby softening the presentation of an alternative view. Such differences suggest that connective choice can subtly influence the tone of disagreement and the writer's evaluative stance toward content.

Yet, current literature in argumentation and writing analytics rarely accounts for these nuances. Most frameworks focus on if a contrastive link is present rather than how it is realized. As research to date has convincingly shown that contrastive connectives are key contributors to cohesion and argument structure, there is a need to move beyond treating *However*, and *While*, with similar meanings, as interchangeable linguistic items. Therefore, this study explores how *However* and *While* are used to maintain coherence and make stances.

3. Methodology

3.1 Data Collection

This pilot study explored how AI chatbots utilize the contrastive connectives *however* and *while* in persuasive debate contexts. The research was set within a computer-mediated discourse environment. Task-oriented chatbots play the role of interlocutors. The purpose of this design was twofold: (1) to provide learners with a controlled but interactive context to engage in persuasive dialogue, and (2) to generate a learner corpus for the systematic analysis of connective use in argumentative discourse.

3.1.1 Participants

Six graduate students participated in the study. Their English proficiency was verified at the CEFR B2 level or higher, which met the departmental requirement for admission. All participants were enrolled at universities in Taiwan. They had no prior experience in formal debate but demonstrated adequate linguistic competence to engage in argumentative tasks. Each participant provided informed consent, and the study was conducted in accordance with ethical research standards.

3.1.2 Chatbot Design

Data were collected using EduACT, a chatbot platform developed by the Department of Computer Science and Information Engineering at National Central University. The platform enables the design of task-oriented conversational agents by specifying agent information, learning topics, task modules, and agent action strategies. Four debate chatbots were created, and that each chatbot were programmed to conduct structured argumentative discussions on the following topics: (a) Should a zoo be built? (b) Should the voting age in Taiwan be set at 18? (c) Should the duration of university semesters be reduced to 16 weeks? (d) Others. Each chatbot was programmed with modular discourse components, including greetings, topic initiation, stance elicitation, argument exchange, rebuttal, wrap-up, and closing. The chatbot language, set to CEFR B1–B2 to ensure comprehensibility for the participants, aimed to make sure participants will focus on employing debat strategies without having a hard time in linguistic choices.

3.1.3 Procedures and Data Collection

Participants were instructed to engage in a debate session with the chatbot for at least 30 minutes. They were informed the topic of choices during the instructions and later decided which debate topic they were going to argue with the chatbot. Also, they were asked to maintain a persuasive stance and respond to the chatbot's prompts, counterarguments, and rebuttals. The chatbot encouraged elaboration and reasoning but provided only limited feedback to ensure the learner's responsibility for sustaining argumentation. All interactions were conducted in English and recorded accordingly by the EduACT platform. The transcripts generated through the task were established as the learner corpus. The dataset comprised a total of 13,482 words of chatbot-generated discourse produced in AI–human debate sessions. This corpus size provides a sufficient basis for examining the syntactic and functional distribution of contrastive connectives as well as their co-occurrence with stance markers and modal verbs. This corpus provides consistent interactional scaffolding while capturing authentic learner responses to argumentative stimuli.

3.2 Data Analysis

The analysis proceeded in several stages. First, all transcripts were cleaned and anonymized in the Excel file and we further examined the frequency and distribution of these connectives to identify their preferred syntactic positions and discourse environments in the AI-generated argumentations. Each occurrence of *While*, and *However* was analyzed within its immediate clause structure to determine the relationship between the preceding clause (X) and the subsequent clause (Y). The term preceding clause refers to the proposition or sentence that appears before the connective, whereas the subsequent clause denotes the proposition or sentence that appears after it. The identification of X and Y was based on syntactic and punctuation cues (e.g., commas, periods) as well as the logical boundaries of meaning within each sentence. In this study, three sentence patterns were analyzed: (1) *While* X, Y, (2) X *while* Y (3) X. *However*, Y.

To provide a coherent analysis of identifying how these connectives contribute to the construction of argumentative stance, this study integrated Hyland's (2005) stance framework with Thetela's (1997) model of evaluative entities. Thetela's model provides the evaluative base whereas Hyland's framework adds strategies. This integration allows each clause or connective-linked proposition to be examined. In this study, content-orientation refers to propositions where connectives link factual or topic-related evaluations that aim to represent external reality with precision rather than negotiate interpersonal meaning. In example (A), *While* connects two factual statements which describe contrasting conditions. The connective links topic-related evaluations of the impact of the economic growth without showing the writer's attitude or reader engagement. Therefore, the proposition is content-oriented and it emphasizes the contrast in reality rather than interpersonal meaning. Writer-orientation denotes propositions where connectives accompany research-oriented evaluations that hedge or distance the writer's personal commitment to claims. In example (B), *however* introduces a qualification that hedges the writer's commitment to the prior claim. The connective signals a shift from assertion to caution with using the modal verb *can* to reflect the writer's assertion of the necessity of further testing. Reader-orientation, by contrast, occurs when connectives introduce evaluations that engage the interlocutor. In example (C), the second proposition introduced by *while* includes the phrase "*we should remember"*, which explicitly invites the reader to participate in the reasoning process, even though in this case the writer uses the modal verb *should*. This usage shows that the writer is acknowledging the reader's role as an active interpreter in the discourse. These criteria provide a systematic approach to categorize stance orientation across propositions.

(A) While economic growth benefits urban populations, rural communities often experience slower development.

(B) *The results appear promising; however, further testing is required before firm conclusions can be drawn.*

(C) *The findings appear encouraging, while we should remember that these outcomes may vary across contexts.*

The data analyzed consist solely of AI-generated texts produced during task-based argumentative interactions with human interlocutors. By isolating the AI's discourse, the analysis can more precisely capture how the system itself constructs cohesion and negotiates stance when responding to argumentative prompts to better facilitate students in enhancing their argumentation skills. The following chapter details the distributional patterns of each connective and discusses their rhetorical functions across the corpus.

4. Results and Discussion

This chapter presents the findings of this study by investigating how the contrastive connectives *however* and *while* were used in AI-human argumentative discourse and this study looked at the AI language use. The discussion focuses on the frequency, distribution, and discourse functions of these two connectives to reveal how they contribute to both cohesion and stance construction in argumentation. The following sections present the quantitative distribution of each connective and discuss how the pattern reflect the AI's cohesive strategies and stance-taking behaviors across argumentative tasks.

4.1 Distributions of Contrastive Connectives

Connective Type	Raw Frequency	Valid Tokens
However	45	40
While (sentence-initial)	68	59
while (sentence-medial)	24	25
Total	137	124

Table 1: Distribution of *while/While* and *However* in the Corpus.

In Table 1, a total of 137 connective tokens were identified through a corpus-based search for the lexical items *however* and *while* across the AI-generated argumentative texts. Each token was manually verified and categorized by its orthographic form (capitalized or lowercase) and sentence position (sentence-initial or sentence-medial). Capitalized forms (e.g., *However*, *While*) were classified as sentence-initial connectives functioning at the discourse level, whereas lowercase forms (e.g., *however*, *while*) were generally categorized as sentence-medial connectives operating within clauses. However, lowercase instances of *while* that followed introductory adverbials such as *Secondly* were also coded as sentence-initial, since they introduced the first clause of the sentence and served to link the current proposition to the preceding argumentative context. No sentence-medial occurrences of *however* were found in the corpus, indicating that the connective was used exclusively in sentence-initial position by the AI system. This consistent capitalization pattern suggests that *however* primarily functioned as a discourse-level contrastive marker, rather than as a clause-internal adverb, in AI-generated argumentation.

After a detailed manual examination, 124 tokens were retained for analysis. Thirteen tokens were excluded because the surrounding discourse provided insufficient contextual evidence to determine their stance orientation—specifically, whether the connected clauses were content-oriented, writer-oriented, or reader-oriented according to Hyland's (2005) framework and Thetela's (1997) model of evaluative entities. Consequently, the final dataset included only those tokens whose stance and cohesive functions could be clearly interpreted within argumentative contexts.

As shown in Table 1, *While* appeared most frequently, accounting for nearly half of all tokens in the corpus, followed by *however* and sentence-medial *while*. The absence of sentence-medial *however* confirms that the AI consistently employed the connective as a sentence-level transition marker, which signals contrast between major argumentative propositions. The relatively high frequency of *While* suggests that the AI often initiated argumentative turns with concessive or contrastive framing. In other words, *While is* used it to structure opposing or qualifying claims.

4.2 Distributional Patterns and Stance Orientation

This section presents the detailed distribution of the 124 valid connective instances and examines how *however* and *while* operate within different stance orientations. The classification followed Hyland's (2005) framework and Thetela's (1997) model of evaluative entities, which functions within a content-oriented (CO) context (factual or topic-related evaluations), writer-oriented (WO) context (reflecting authorial stance), or reader-oriented (RO) context (inviting reader engagement). This analysis allows for a more nuanced interpretation of how the AI constructs logical relations and expresses evaluative positioning through contrastive connectives in argumentative interactions.

The distributional analysis first focused on the preceding and subsequent clauses linked by each connective, revealing how these relations contribute to local cohesion and stance realization. Quantitative results were then interpreted in light of the connective's rhetorical role in order to show how AI-generated argumentation builds coherence and authority. Table 2 summarizes the combined distribution of *however* and *while* across stance categories in both clause positions.

Connective Type	Preceding Clause			Subsequent Clause		
	CO	RO	WO	CO	RO	WO
However	25	10	5	1	2	37
While (sentence-initial)	56	0	3	1	0	58
while (sentence-medial)	12	0	13	22	1	2
Total	93	10	21	24	3	97

Table 2: Combined Distribution of *while/While* and *However* by Orientation in Preceding and Following Clauses (N = 124).

4.2.1 Functional tendencies in the preceding clause

In the preceding clause (X), the connectives primarily serve content-organizing purposes, which links propositions through logical, contrastive, or concessive relations that construct the interactive arguments. Among the connectives, *While* (capitalized) frequently introduces a contextual or contrastive/ concession discourse against which the subsequent clause develops the main stance.

(1) CO: …*While visiting companies can offer firsthand exposure, it is important to note that classroom education provides a foundation of knowledge,…*

In the excerpt (1), the preceding clause *"While visiting companies can offer firsthand exposure"* presents a contextual concession that acknowledges the potential benefits of an opposing view before the AI continues its main argument in the following clause. The connective *While* here introduces a background condition that appears to align with the opponent's reasoning but immediately contrasts it with a stronger counter-assertion, which is "classroom education provides a foundation of knowledge." This structural move reflects the concessive/adversative function of *While* as defined by Halliday and Hasan (1976), where the initial clause serves to moderate or anticipate disagreement while preserving logical continuity. Functionally, this use of *While* demonstrates how the AI organizes argumentative discourse by balancing acknowledgment and rebuttal and further creates a smooth transition between opposing propositions without disrupting overall cohesion.

Similarly, *However* appearing in preceding clauses frequently introduces a counter-claim or an alternative perspective that contrasts with the immediately preceding proposition.

(2) CO: ... *that they play a crucial role in conservation efforts by protecting endangered species and educating the public about wildlife conservation. However, critics argue that these goals can be achieved through alternative means such as wildlife sanctuaries or nature reserves without…*

In excerpt (2), *However* marks a discourse shift from an affirmative stance to an opposing viewpoint, guiding readers through a transition between conflicting propositions. The connective operates primarily as a content-oriented meta-discourse marker, organizing propositional relations rather than conveying personal evaluation. According to Hyland's (2005) framework, such usage indicates that the AI's focus remains on maintaining logical and cohesive contrast, emphasizing informational balance rather than interpersonal engagement. Through this function, *However* intensifies the structural integrity of the argument while delineating competing claims within the same topical frame.

(3) CO: …*lowering the voting age can provide young people with an opportunity for political participation from an earlier stage of life while*

also fostering civic engagement among youth. In conclusion: Lowering the voting age allows young people aged 18 or above who…

By contrast, the sentence-medial *while* often operates within the sentence to connect simultaneous or concessive conditions. In excerpt (3), *while* functions as an intrasentential connector, linking two clauses that share thematic continuity and occur in parallel. The connective signals syntactic subordination, which emphasizes propositional linkage rather than a shift in stance. This corresponds to Blühdorn's (2008) description of subordination in syntax and discourse, where *while* does not introduce a new argument but rather integrates a supporting or co-occurring condition within the same argumentative frame. Thus, the clause preceding *while* contributes to the content-oriented organization of ideas, maintaining cohesion through logical simultaneity instead of contrastive opposition.

A smaller portion of preceding clauses (ten cases) involved *However* functioning as an attentional cue that directs the reader's focus and signals a forthcoming shift in reasoning. According to Hyland's (2005) framework, such use represents a reader-orientation. The connective helps readers follow the logical flow of argumentation rather than expressing the writer's personal attitude.

(4) RO: *…I believe your concern, XXX.* <u>*Learning a foreign language does require more time and practice compared to other subjects.*</u> *However, I believe that reducing the total number of weeks in the university to 16 can still be beneficial for language…*

In excerpt (4), *However* performs a reader-guiding function, which helps the reader to anticipate a contrastive move while maintaining textual coherence. This function reflects discourse management, that is, language is used to structure understanding and facilitate reader processing. This shows that the AI's connective use prioritizes textual clarity and reader comprehension over interpersonal engagement.

Meanwhile, writer-oriented preceding clauses (21 instances) exhibit explicit stance marking. It is often realized through evaluative or modal expressions embedded within *while*-clauses. These linguistic elements project judgment, intention, or obligation, thereby signaling the writer's evaluative control over the discourse rather than neutral description.

(5) WO: *…*<u>*This way, students can engage in meaningful discussions*</u> *while also gaining practical experience outside the classroom. …*

In excerpt (5), the modal "*can*" expresses possibility and positive potential, which frames the action as both attainable and desirable. The evaluative adjective "meaningful" further marks the writer's approval to project value judgment within the subordinate *while*-clause. These linguistic cues position the AI as asserting that experiential learning is beneficial, thereby intensifying an affirmative stance.

(6) WO: *…*<u>*we ensure that these benefits are maximized*</u> *while minimizing any negative impacts on animal welfare or environmental concerns…*

Similarly, in excerpt (6), the verbs "ensure" and "maximize" convey intentionality and moral evaluation; while the present participle "promoting" embodies an evaluative implication of social good. Across these examples, the embedded *while*-clauses function as stance-bearing extensions instead of neutral elaborations. This confirms Hyland's (2005) view that stance is often realized lexically and grammatically through evaluative or modalized language, which in turn aligns with Thetela's (1997) model of writer-oriented evaluation. That is, authors project their beliefs and commitments explicitly in text.

In sum, *While* and *However* in the preceding position primarily function as textual cohesive devices, which organizes propositional flow and marks transitions between argumentative turns. Their roles are largely structural rather than interpersonal. It supports the logical development of ideas and signals shifts in argumentative direction rather than projecting the writer's personal stance. Through these connectives, the AI manages discourse coherence and establishes clear rhetorical boundaries across turns, and further maintains smooth progression within the argumentative sequence.

4.2.2 Functional tendencies in the subsequent clause

While the preceding clauses primarily serve to organize propositional content and establish logical transitions, the subsequent clauses (Y) reveal how the AI develops stance and evaluation in response to the ideas introduced earlier. Table 2 shows the dominance of *writer-oriented* functions (97/124, 78.2%) across the connectives. In this proposition, connectives often signal a shift from exposition to evaluation, which transforms informational content into an explicit argumentative stance. The subsequent clause thus becomes the site where writer-oriented expression emerges most strongly, which is frequently realized through evaluative lexis, modal verbs, or contrastive assertions that consolidate the AI's position. By examining these tendencies, this section demonstrates how the AI constructs cohesion while simultaneously intensifying its argumentation within the discourse sequence.

However overwhelmingly performs this evaluative function, with 37 of its 40 subsequent clauses to express the writer's viewpoint, qualification, or interpretive stance.

(7) WO: …*It is indeed a valuable experience that can provide them with a different perspective on the industry or specific companies. However, <u>I would like to propose that classroom education and real-world experiences are not mutually exclusive but rather complementary</u>…*

In excert (7), the connective *However* marks a clear shift from factual description to interpretive evaluation. The preceding clause presents a neutral, content-oriented observation, while the subsequent clause, introduced by *However,* reframes the discussion through a writer-oriented assertion, "…I would like to propose…*"*. This illustrates how *however* shifts from a cohesive organzier to a stance-marking device. It allows the AI to project argumentative authority and evaluation in the argumentation. The connective thus performs a dual role, which is to maintain logical contrast while simultaneously emphasizes authorial judgment. This rhetorical function aligns with Hyland's (2005) notion of stance as authorial presence, in which meta-discourse markers mediate the writer's epistemic and attitudinal positioning.

Similarly, *While* (capitalized) frequently introduces a subordinate clause followed by a writer-oriented main clause that conveys the AI's principal claim or evaluation.

(8) WO: *…In conclusion, while 16 weeks may seem limited at first glance, <u>it can actually promote focused learning, active engagement with the material, independent</u>…*

In excerpt (8), the connective *While* establishes a concessive background (…16 weeks may seem limited…) that legitimizes opposing expectations before transitioning to the main evaluative statement (…it can actually promote focused learning…). This structure demonstrates how *While* contributes to argumentative coherence through concession and rebuttal, which allows the writer to acknowledge potential counterarguments while intensifying their own position. The pattern reflects Hyland's (2005) description of writer-oriented metadiscourse, where concessive connectives facilitate stance construction by managing the interplay between alternative perspectives or evaluation.

The sentence-medial *while* (22 content-oriented and 2 writer-oriented instances) in the subsequent clause shows a more distinctive stance orientation. This distribution indicates that the connective primarily supports local semantic cohesion rather than projecting evaluation or authorial stance.

(9) CO: *…Ultimately, finding an optimal balance between instructional time and assessments is essential to ensure quality education while <u>considering factors unique to each educational system</u>…*

In excerpt (9), *while* integrates a proceding clause that specifies a contextual condition (…considering factors unique to each educational system…), marked as content-oriented. This type of *while*-clause fulfills what Blühdorn (2008) terms syntactic subordination, which foregrounds propositional linkage and logical dependency rather than meta-discursive or evaluative commentary.

(10) WO: *…lowering the voting age can provide young people with an opportunity for political participation from an earlier stage of life while <u>also fostering civic engagement among youth.</u>" „"I believe that in Taiwan's law, individuals are considered full adults at…*

By contrast, in excerpt (10), *while* connects two propositions that express the writer's stance toward youth empowerment. The evaluative element is realized through the adjective "civic" and the lexical choice "engagement", both of which carry positive connotations of active citizenship. Here, *while* functions as a stance-marking connector, intensifying the writer's approval of the underlying social goal. Despite its limited frequency, this pattern demonstrates that the AI occasionally deploys *while* to strengthen positive evaluation and assert authorial perspective.

Only three instances in the subsequent clause were reader-oriented. It serves as the clarification that invite mild reader reflection without direct interaction.

(11) RO: *...natural habitats. By implementing these improvements, zoos have the potential to become centers for education, conservation, and ethical animal care while* <u>*still allowing people to learn about and appreciate wildlife.*</u> *Now it's your turn! How would you respond?"...*

(12) RO: *...foreign countries like the UK and the USA. It's true that assessments can be scheduled outside of those 12 weeks. However,* <u>*it's important to consider that different educational systems have different approaches to assessments*</u>*. In Taiwan's education system,...*

In excerpt (11) and (12), the *while*-clauses serve as reader-oriented cues that gently acknowledges the reader's perspective (*allowing people to learn/ it's important to consider*). However, such little instances support Thetela's (1997) observation that academic writers, including AI, tend to maintain interpretive control rather than engage readers directly. This orientation intensifies the AI's preference for informational cohesion.

4.2.3 Comparative Interpretation

The combined findings reveal a functional progression across clause positions. In other words, the orientations move from content orientation in the preceding clause to writer orientation in the subsequent clause. This shift underscores how connectives enable the AI to shift from logical organization to evaluative stance. It guides readers through both the propositional structure of the argument. The pattern reflects an underlying rhetorical design: (1) The proceding clause establishes informational grounding, (2) The subsequent clause asserts perspective and authorial control.

Connectives (*While* (sentence-initial) and *However*) initially structure logical or concessive/contrastive relationships (mostly content-oriented) in the preceding clause (X) and then progress toward writer-oriented evaluation or claim articulation in the subsequent clause (Y) (see excerpt (13)). It exemplifies how the AI constructs coherence through both semantic contrast and stance advancement.

(13) CO/ WO: *...While* <u>*independent learning is important,*</u> <u>*It is also crucial for students to engage in structured classroom settings*</u>*. University classes provide...*

In contrast, the sentence-medial *while* primarily functions as a semantic connector rather than a meta-discursive device. It maintains intra-sentential cohesion without overt stance marking. This aligns with Blühdorn's (2008) argument that syntactic subordination does not necessarily entail discourse-level hierarchy. That is, a subordinating conjunction like *while* may grammatically link clauses without indicating a rhetorical or evaluative relation. The connective thus sustains textual cohesion but remains limited in its capacity to project interpretation or evaluation. This evidence supports Hyland's (2005) claim that meta-discourse serves as a dual resource, which is organizational in structuring propositional flow and interpersonal in expressing stance and authorial control.

Type	Dominant Orientation Pattern	Preceding Clause/Subsequent Clause
However	CO → WO	Establish factual or contextual premises/ Present evaluation, claim, or counter-claim
While (sentence-initial)	CO → WO	Providing contextual or concessive background information/ Expressing the main stance or conclusion
while (sentence-medial)	CO → CO	Describing states or conditions/ Maintaining topical continuity or logical simultaneity

Table 3: Comparison of Stance Orientations and Functional Roles of *However* and *While*

Overall, the distribution reveals a functional progression across positions. *However* and sentence-initial *While* typically shift from content orientation in the preceding clause to writer orientation in the subsequent clause. This shows how the AI shifts from describing facts to expressing evaluation. In contrast, sentence-medial *while* remains content-oriented. It emphasizes cohesion and logical continuity rather than stance construction (see Table 3). The findings therefore confirm that the AI's connective use not only organizes textual relations but also simulates the evaluative progression characteristic of human academic discourse.

5. Conclusion

This study examined how the contrastive connectives *however* and *while* function in the AI–human argumentative discourse. Drawing on Hyland's (2005) stance framework and Thetela's (1997) model of evaluative entities, the analysis revealed a consistent rhetorical progression across clauses: from content-oriented organization in preceding clauses to writer-oriented evaluation in subsequent ones. The findings suggest that the AI constructs argumentation through a two-step rhetorical strategy: (1) to establish logical balance, (2) to assert stance. This shows the cohesive and persuasive characteristic of academic writing. These results extend connective research by showing that *however* and *while* perform distinct discourse functions in constructing coherence and projecting stance.

Acknowledgement

The authors would like to thank all the reviewers whose comments helped improve this paper. This research was partially supported by the National Science and Technology Council under Project Number: 113-2410-H-004-064-MY2.

References

Abdi, O. (2021). The use of discourse connectives in the written academic discourse of students majoring in Arabic and their peers majoring in English. *Arab Journal of Applied Linguistics*, *6*(01), p. 32-59.

Blühdorn, H. (2008). *Subordination and coordination in syntax, semantics and discourse: Evidence from the study of connectives*. In C. Fabricius-Hansen & W. Ramm (Eds.), 'Subordination 'versus 'Coordination 'in Sentence and Text. Amsterdam: John Benjamins. p. 59–85.

Eckle-Kohler, J., Kluge, R., & Gurevych, I. (2015, September). On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2236-2242).

Felder, E. (2015). Lexik und Grammatik der Agonalität in der linguistischen Diskursanalyse. *Diskurs–interdisziplinär. Zugänge, Gegenstände, Perspektiven*, 87-121.

Fraser, B. (1999). What are discourse markers?. *Journal of pragmatics*, *31*(7), p. 931-952.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Routledge

Kamalski, J., Lentz, L., & Sanders, T. (2006). Effects of coherence marking on the comprehension and appraisal of discourse. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 28, No. 28). p. 1575-1580.

Maamuujav, U. (2025). Relations of linguistic features, rhetorical moves, and writing quality in academic writing of secondary students. *International Journal of Educational Research Open*, *9*, 100505. p. 1-14.

Rehbein, I. (2019, August). On the role of discourse relations in persuasive texts. In *Proceedings of the 13th Linguistic Annotation Workshop*. pp. 144-154.

Schiffrin, D. (1987). *Discourse marker*s. Cambridge University Press.

Stab, C., & Gurevych, I. (2014, October). Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 46-56).

Thetela, P. (1997). Evaluated entities and parameters of value in academic research articles. *English for specific purposes*, *16*(2), 101-118.

Quantum Perspectives on Persuasive Language in AI-Generated News: A QNLP-Based Analysis

Jung-Hua Liu
Department of Communication, National Chung Cheng University / No. 168, Sec. 1, University Road, Minxiong Township, Chiayi County 621301, Taiwan
emileliu@ccu.edu.tw

Abstract

This study applies quantum natural language processing (QNLP) to 298 Chinese AI-generated YouTube news articles. Using IBM Qiskit, this study reveal multi-framing narratives with high frame competition but low conflict. Headlines employ emotion, content stays neutral or positive, showing strategic ambiguity. QNLP metrics highlight persuasive tactics and implications for communication theory and AI ethics.

Keywords: Quantum NLP, persuasive language, framing, agenda-setting, strategic ambiguity

1 Introduction

Generative AI now produces news articles, raising questions about authenticity, framing, and persuasion (Paviour, 2025). Building on Lippmann's (2017) idea that news constructs a "pseudo-environment," Yin & Liu (2025) and Reubold, and Campbell (2023) note that AI-driven journalism transfers gatekeeping from editors to algorithms. Classic theories of agenda-setting and framing remain relevant: while human editors once selected topics and angles, AI systems may inherit training-data biases or create new emphases (Mehrab et al., 2021; de-Lima-Santos & Jamil, 2024; Kuku et al., 2025; Singh & Ngu, 2025). Studies show AI news can differ in style and tone from human reporting but is not necessarily more biased (Nah et al., 2024; Sui, 2025). Recent studies on AI-generated discourse reveal that persuasion in digital communication extends far beyond surface-level sentiment or credibility measures. Goldstein et al. (2024) demonstrate that GPT-3 can generate propaganda nearly as persuasive as authentic state-backed content, particularly when human curators refine or select output. This finding underscores that persuasive efficacy in AI-generated text emerges not solely from factual accuracy but from rhetorical coherence, emotional framing, and contextual adaptability. Similarly, Pazzaglia et al. (2025) show that fine-tuned large language models reproduce polarized ideological rhetoric with high credibility and emotional resonance. Their model's outputs were rated as both "provocative" and "human-like," suggesting that persuasive force arises from the capacity of language models to reproduce rhetorical alignment - a blending of ideological tone, emotional activation, and discursive context. Meanwhile, Al Giffari and Dermawan (2025) reveal through comparative rhetorical analysis that AI-generated religious messages, though formally coherent and citation-driven, lack the ethos, pathos, and kairotic timing that human preachers use to achieve moral and emotional persuasion.

These works indicate that persuasion in AI discourse depends not only on propositional content but on the quantum-like coexistence of multiple interpretive frames -logical, emotional, and ethical - that audiences navigate dynamically. Quantum Natural Language Processing (QNLP) provides a theoretical and computational framework for representing this multidimensional interplay. By encoding textual meaning as quantum states, QNLP models semantic superposition (simultaneous coexistence of conflicting frames), entanglement (interdependence among linguistic and contextual cues), and measurement collapse (resolution of ambiguity through interpretation). In persuasion analysis, these quantum phenomena map onto how readers oscillate between alternative framings or emotional cues before forming conviction—analogous to the probabilistic collapse of a quantum system upon observation. This approach captures what traditional NLP misses: that persuasive communication operates through contextual interference patterns among emotional tone, narrative perspective, and cultural resonance.

QNLP thus formalizes persuasion as an emergent property of narrative context rather than as a unidirectional rhetorical act. It illuminates how AI-generated news or propaganda can appear simultaneously neutral, credible, and manipulative - precisely because its semantic space allows multiple persuasive potentials to coexist until interpretively resolved by the audience. Through this lens, QNLP bridges computational linguistics and communication theory, offering a post-classical model for analyzing how machine-produced narratives shape belief, trust, and ideological alignment in the quantum field of discourse. Findings highlight strategic ambiguity,

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 358–368
November 20-22, 2025 ©2025 Association for Computational Linguistics

emotional framing, and expanded agenda breadth as persuasive features. QNLP thus bridges communication theory and quantum semantics, offering new tools for detecting subtle persuasive strategies in AI-generated content.

2 Literature Review

2.1 Persuasion through Agenda-Setting, Framing, and Rhetorical Strategies

Persuasion in news discourse has long been theorized through the intertwined mechanisms of agenda-setting, framing, and rhetorical strategy. These perspectives, while often treated separately, all explain how media shape public attitudes not by direct argumentation but by structuring attention, interpretation, and affective response—the key ingredients of persuasion.

Agenda-setting theory (McCombs & Shaw, 1972) shows that the persuasive force of news lies in its power to prioritize certain topics over others, implicitly signaling their importance. At the first level, issue salience determines what the public thinks about; at the second level, attribute salience determines how they think about it. For example, emphasizing unemployment statistics rather than individual hardships frames the issue as technical rather than moral, guiding public concern and policy preferences. This process is persuasive because it conditions cognitive accessibility: repeated exposure elevates certain issues in collective awareness, creating perceived consensus and urgency.

Framing theory deepens this account by showing that persuasion occurs through selection and emphasis. Entman (1993) defined framing as the act of selecting aspects of perceived reality to make them more salient, thereby promoting specific problem definitions, causal interpretations, moral evaluations, and policy recommendations. Frames thus operate as interpretive templates that steer reasoning. A protest described as a "law-and-order problem" activates threat and control schemas, whereas the same event framed as a "civil-rights struggle" evokes empathy and justice. In both cases, framing does not merely present facts—it organizes meaning in ways that predispose audiences toward particular attitudes or actions.

Rhetorical strategies complete the persuasive triad by illuminating how linguistic and stylistic choices translate cognitive framing into affective engagement. Classical rhetoric's ethos, pathos, and logos correspond to credibility, emotion, and logic—the dimensions that sustain belief formation. Even under norms of journalistic objectivity, subtle rhetorical cues such as evaluative adjectives, quotation patterns, or metaphoric phrasing convey stance and invite alignment. Ceccarelli's (1998) concept of strategic ambiguity further explains how persuasion can arise from texts that support multiple plausible interpretations: ambiguity minimizes resistance by allowing diverse audiences to read agreement into the same message. Thus, persuasion in journalism is often implicit, operating through agenda prominence (what to think about), framing (how to think about it), and rhetorical form (how to feel about it). These mechanisms collectively blur the boundary between informing and influencing, creating an ecology of persuasion that relies on selection, emphasis, and affect rather than overt argumentation.

2.2 Persuasion in the Age of AI and NLP

The rise of artificial intelligence (AI) in journalism, often referred to as automated or robot journalism, has intensified scholarly attention to persuasion's algorithmic dimensions. During the 2010s, outlets such as Reuters, the Associated Press, and The New York Times adopted rule-based generators for financial reports and sports summaries (Carlson, 2018; Diakopoulo, 2019; The Newsreel Project Consortium, 2021). By the 2020s, large language models (LLMs) like OpenAI's GPT series enabled generative systems to produce multi-paragraph narratives that mimic human style and rhetorical nuance.

Recent studies highlight both opportunities and ethical challenges. A systematic review by Ioscote et al. (2024) notes that automation improves efficiency but introduces opacity and potential bias. Graefe (2016) found that algorithmic news was perceived as competent but emotionally flat, while Nah et al. (2024) observed that AI-generated stories differ in tone and coherence yet are not necessarily more biased. Nonetheless, AI's capacity to synthesize persuasive patterns from vast corpora gives it unprecedented influence over public cognition. Goldstein et al. (2024) showed that GPT-3-generated propaganda can be nearly as persuasive as human-written material, especially when curated by humans. Pazzaglia et al. (2025) found that fine-tuned LLMs replicate polarized discourse with rhetorical realism, while Al Giffari and Dermawan (2025) demonstrated that AI reproduces logical appeals but lacks the adaptive ethos and emotional depth of human persuasion. These findings converge on one point: AI's persuasive power lies in its ability to simulate the *agenda-setting and framing patterns* that shape interpretive hierarchies in human journalism.

Even absent malicious intent, AI systems reproduce persuasive conventions—issue prioritization, emotional tone, narrative balance, and ambiguity—because these features are embedded in their training data. Conventional NLP tools such as sentiment analysis or topic modeling can capture tone and frequency but cannot fully represent how frames interact or compete within a narrative. Similarly, propaganda-detection systems focus on lexical signals but overlook the contextual superpositions that make messages persuasive across ideological lines.

QNLP provides a post-classical approach to this challenge. By encoding text as quantum states, QNLP models superposition (simultaneous activation of multiple frames), entanglement (interdependence among topics, emotions, and rhetorical cues), and measurement collapse (resolution of interpretive ambiguity during audience reception). These quantum concepts parallel how persuasion unfolds in narrative contexts: audiences oscillate between competing frames and affective interpretations before settling on belief or skepticism. QNLP thus allows researchers to formalize and visualize the non-linear, context-dependent nature of persuasion—how agenda-setting, framing, and rhetoric operate together to construct probabilistic meaning fields rather than fixed messages.

Classical theories reveal that persuasion in journalism emerges from the coordination of attention (agenda-setting), interpretation (framing), and affect (rhetoric). In the AI era, these mechanisms are not only replicated but amplified by generative systems capable of producing multi-frame, strategically ambiguous narratives at scale. QNLP offers a novel alternative. By encoding texts as quantum states, QNLP enables analysis of overlapping meanings, frame superpositions, and narrative entanglements. This study is among the first to employ QNLP to examine persuasive dynamics in AI-generated news, particularly focusing on frame competition, ambiguity, and agenda breadth (Wazni et al., 2024; Widdows et al., 2024). Integrating QNLP into this analytical framework offers a powerful means to decode the entangled semantics and contextual fluidity of persuasion in AI-generated news.

3 Methodology

3.1 Dataset and Corpus Preparation

This study generated and analyzed a dataset of 298 GPT-4o-generated Chinese-language news articles obtained from a YouTube channel that produces automated news videos. These videos drew content from Yahoo! News across domains such as politics, economy, technology, and society. Each news item contained three textual components:

- **News Title**: Averaging 16 Chinese characters, titles conveyed the story's core point or a teaser. They were designed to attract attention, often using emotion or framing (e.g., *"Tech CEO Promises Reform Amid Crisis"*).
- **Video Dialogue (Transcript)**: Averaging 334 characters, dialogues resembled talk-show or multi-speaker formats, simulating anchors and guests. This style incorporated multiple perspectives, quotes, and facts.
- **Video Description**: Averaging 256 characters, descriptions summarized key points and context, functioning as concise press-release style overviews.

Together, these three layers provided a multi-tiered discourse structure: headlines framed events with emotional hooks, dialogues expanded perspectives through conversation, and descriptions offered neutral summaries. This layering enabled analysis of persuasive strategies at different textual levels.

The dataset covered diverse topics, ensuring generalizable findings beyond a single domain. While modest in size (n=298), the corpus allowed meaningful quantitative analysis while remaining computationally manageable.

To establish a baseline and strengthen empirical grounding, the QNLP pipeline was applied to a comparative dataset of professionally written news from Taiwan's Central News Agency (CNA), the nation's official wire service. The dataset comprised 20 paired samples of news titles and full articles, all published in 2020, thereby ensuring that the material predated the widespread adoption of AI-assisted or AI-generated writing. This corpus served as a human-authored benchmark for assessing whether the distinctive characteristics observed in AI-generated texts - such as high frame competition, low conflict, and mild positivity - are unique to algorithmic generation or instead reflect broader conventions of traditional journalistic discourse.

Text Preprocessing: Since Chinese lacks spaces, word segmentation was essential. The Jieba tool was used to split text into lexemes (e.g., "人工智慧" as *artificial intelligence* rather than "人工" + "智慧"). After segmentation, standard cleaning included normalizing full-width to half-width characters, ensuring UTF-8 encoding, and removing non-textual artifacts. Stopword removal was not applied, as function words carry meaning important for QNLP. All analysis was conducted in Chinese. Each component (title, dialogue, description) was analyzed separately and comparatively to reveal differences in tone, framing, and entropy.

3.2 Quantum NLP Encoding with Qiskit

The Quantum Natural Language Processing (QNLP) pipeline was implemented using IBM's Qiskit. Following the DisCoCat model (Coecke et al., 2010; Meichanetzidis et al., 2020), text was encoded as quantum states to represent semantic and narrative features.

The Jieba library performs segmentation and part-of-speech tagging. Each segmented and POS-tagged Chinese word is encoded as a quantum state |ψ_word⟩, where its grammatical role determines the number of qubits used and how they interact within the circuit. Mapped tags follow the DisCoCat (Categorical Compositional Distributional) model types (see Table 1). Each part-of-speech category is assigned a **pregroup type** which is mapped to a **vector-space representation** $T_y(-)$under the strong monoidal functor F:Pregroup → FVect. The notation $T_y(n)$denotes the vector space corresponding to the noun type $n$under F. The tensor product symbol (⊗) indicates the compositional combination of vector spaces (or linear maps) to represent joint meaning and grammatical interaction in the DisCoCat framework:

POS	Category	DisCoCat Type
n, nr, nt	Noun (N)	$T_y(n)$
v, vn	Verb (V)	$T_y(n)^r \otimes T_y(s) \otimes T_y(n)^l$
a	Adjective (A)	$T_y(n) \otimes T_y(n)^l$
d	Adverb (D)	$T_y(s) \otimes T_y(s)^l$
p	Preposition (P)	$T_y(n)^r \otimes T_y(n) \otimes T_y(n)^l$

Table 1: Pregroup Type → Vector-Space Mapping.

Each qubit acts as a semantic container that can represent multiple potential meanings simultaneously just as a word such as "改革" (reform) may convey both positive and critical implications depending on context. Unlike classical bits that exist only as 0 or 1, a qubit can occupy a superposition
$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, representing a weighted combination of interpretive possibilities. In the QNLP model, each part of speech (POS) corresponds to a grammatical type that specifies how its meaning composes with others:

Noun (N): a single-qubit subsystem $T_y\,(n)$ representing an entity.

Verb (V): a composite subsystem $T_y\,(n)^r \otimes T_y\,(s) \otimes T_y\,(n)^l$ that links two noun qubits (subject and object) through entanglement.

Adjective (A): a two-qubit structure $T_y\,(n) \otimes T_y\,(n)^l$ that modifies a noun.

Adverb (D): a two-qubit structure $T_y\,(s) \otimes T_y\,(s)^l$ that modifies a verb or clause.

Preposition (P): a three-qubit subsystem $T_y\,(n)^r \otimes T_y\,(n) \otimes T_y\,(n)^l$ that introduces relational meaning.

There is an example as the below one,

Sentence:「麥當勞性侵案後改革 董事長發聲承諾改善」

(*Màidāngláo xìngqīn àn hòu gǎigé dǒngshìzhǎng fāshēng chéngnuò gǎishàn*; *After the McDonald's sexual assault case, the chairman spoke out and promised reform.*)

Segmentation and POS tagging:

Output:

[("麥當勞 (Màidāngláo, McDonald's)", 'nt'),
("性侵 (xìngqīn, sexual assault)", 'n'),
("案 (àn, case)", 'n'),
("後 (hòu, after)", 'f'),
("改革 (gǎigé, reform)", 'v'),
("董事長 (dǒngshìzhǎng, chairman)", 'n'),
("發聲 (fāshēng, to speak out)", 'v'),
("承諾 (chéngnuò, to promise)", 'v'),
("改善 (gǎishàn, to improve)", 'v')]

Mapped to grammatical categories following DisCoCat: [N, N, N, F, V, N, V, V, V].

Here:

- **N (noun)** = organization / entity
- **V (verb)** = action or statement
- **F (function)** = adverbial or time marker ("後", after)

The base model uses eight qubits representing major Chinese grammatical categories (noun,

verb, adjective, adverb, preposition, pronoun, conjunction, other). Additional qubits (up to four) are allocated proportionally to the number of unique part-of-speech tags and compositional transitions, ensuring that texts with richer syntactic variation yield more entangled circuits. The algorithm means the more **unique POS tags** (diversity of grammar) a sentence contains, And the more **category transitions** (e.g., N→V→N→F→V) occur,
→ The higher the **compositional complexity**, and thus, more qubits are added.

These qubits don't represent specific parts of speech — they capture **semantic entanglement patterns** such as:

- **Noun–Verb entanglement:** subject–predicate dependencies.
- **Adjective–Noun entanglement:** modification dependencies.
- **Temporal–Action coupling:** time or cause-effect encoding.

For example, in the above case, "麥當勞性侵案後改革 董事長發聲承諾改善", there are 9 tokens**,** 4 major POS categories (N, V, F, A), and multiple inter-category transitions:
N → N → N → F → V → N → V → V → V
yielding a compositional complexity value high enough to allocate 4 extra qubits (Table 2).

Qubit	Linguistic Role	Example Representation
q_0	Noun (Subject)	麥當勞 (McDonald's)
q_1	Verb (Action)	改革 (reform)
q_2	Function / Modifier	後 (after)
q_3	Complement Noun	董事長 (chairman)
q_4	Adjective / Evaluation	良好 (good)
q_5	Adverb / Tone	積極 (actively)
q_6	Contextual Frame	政治 / 經濟 (political/economic)
q_7	Rhetorical Mode	Hopeful / Critical tone

Table 2: Qubit Type

This 8-qubit configuration allows this model to:

- Encode **frame superposition** (multiple meanings or framings coexisting).
- Maintain **semantic entanglement** (how grammatical roles affect each other).
- Simulate **interpretive collapse** (when a reader resolves ambiguity).

4 base qubits = structural grammar**,**
4 extra qubits = higher-order semantic entanglement. Together, they form a full 8-qubit quantum linguistic state**:**

$$|\psi\rangle = \Sigma_{i=0}^{2^8-1} \alpha_i \mid i\rangle$$

where each amplitude α_i corresponds to a possible interpretive configuration of the sentence. Each of the 256 possible configurations (from |00000000⟩ to |11111111⟩) represents a distinct combination of meanings. Each amplitude α_i captures the weight or probability of that interpretation. When measured (interpreted by a reader), the sentence collapses into one dominant interpretation. If many amplitudes are large, the sentence is ambiguous with multiple frames; if one dominates, the meaning is singular.

In the Qiskit implementation:

- **Hadamard (H)** gates initialize the emotional subsystem into superposition.
- **Rotation (R_Y)** gates encode each quantum weight as a rotation angle.
- **Controlled-NOT (CX)** and **Controlled-RZ (CRZ)** gates introduce entanglement when both positive and negative cues occur, simulating **emotional interference** between coexisting sentiments.

this design allows the circuit to capture complex emotional polarity interactions, e.g., optimism and anxiety expressed simultaneously within reform narratives.

Each sentence is converted to a quantum circuit through three main steps

1. **Initialization:** All qubits begin in superposition states via Hadamard gates: $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$, representing interpretive openness.
2. **Category-Specific Rotations:** Each grammatical category applies rotation gates proportional to its frequency and semantic role. Rotation about the Y-axis, $R_Y(\theta)$, encodes meaning amplitude; phase rotations (R_Z) introduce semantic distinctions.

3. **Entanglement:** CNOT (CX) and controlled rotation (CRY) gates encode syntactic dependencies such as noun–verb or adjective–noun relationships.

Here, nouns form base qubits, verbs are modeled as multi-qubit subsystems, and grammatical dependencies such as noun–verb or adjective–noun pairs are represented through entanglement.

3.3 Quantum Metric Definitions

Defined as the normalized von Neumann entropy of the article's density matrix (Widdows, et al., 2024; Agostino et al., 2025):

$$C = \frac{-\mathrm{Tr}(\rho \log_2 \rho)}{\log_2 N}$$

where:

- **ρ** is the *density matrix*, which represents all possible meanings or frames encoded in the text's quantum state. It is constructed as $\rho = |\psi\rangle\langle\psi|$, the outer product of the statevector with itself.
- **Tr(ρ log₂ ρ)** means taking the *trace* (sum of diagonal elements) of the matrix after applying a logarithm. This operation computes the weighted average uncertainty of the entire meaning distribution.
- **−Tr(ρ log₂ ρ)** gives the *von Neumann entropy*, a measure of how mixed or diverse the meanings are.
- **Dividing by log₂ N** (where *N* is the number of possible interpretive frames) normalizes the result between 0 and 1.

If $C \approx 1$, the text contains multiple equally active frames (e.g., political, moral, and economic frames appearing together).
If $C \approx 0$, one frame dominates and the article has a single clear angle.
Thus, *C* quantifies how much interpretive "competition" exists in the text's meaning structure.

Von Neumann Entropy

$$S(\rho) = -\mathrm{Tr}(\rho \log_2 \rho)$$

where:

- **S(ρ)** is the *von Neumann entropy*, quantifying semantic uncertainty or agenda diversity.
- **ρ** is the density matrix of the encoded quantum-linguistic state.
- **Tr** denotes the trace, and **log₂** computes information in bits.

Higher entropy values (approaching 0.9) imply broad topical or interpretive diversity; lower values indicate focused discourse.

3.4 Quantum-weighted Sentiment

Emotional tone analysis was implemented through a **heuristic quantum-weighted lexicon**, where each emotional token is assigned a polarity weight representing its affective intensity in the range 0.65–0.95. Because existing Chinese sentiment benchmarks do not support the QNLP encoding framework, this study constructed a self-calibrated emotional lexicon based on high-frequency evaluative terms observed in the dataset.
The complete emotion arrays used for analysis are listed below.

These weights serve as quantum amplitudes reflecting how strongly each emotional concept contributes to the sentence's overall affective state before normalization.

After segmentation and POS tagging, each emotional word is matched with its corresponding weight w_i. Sentence-level emotional intensity is calculated as:

$$I_{\text{emotion}} = \Sigma_i (w_i \times \text{count}_i)) / \text{total\_words}$$

The resulting intensity is then mapped to a rotation parameter for quantum circuit encoding:

$$\theta_i = \pi \times I_{emotion}.$$

Thus, a word with weight $w_i = 0.78$ produces a rotation $R_Y(0.78\pi)$, generating the corresponding emotional amplitude in the quantum state.

Syntactic patterns further refine the emotional amplitude:
• Active-voice markers ("主動 zhǔdòng – active / initiative", "積極 jījí – positive / energetic", "推動 tuīdòng – to promote / to drive forward") add up to +0.10 to strengthen positive orientation.
• Future markers ("將 jiāng – will / shall", "會 huì – will / be likely to", "計劃 jìhuà – plan / project") add up to +0.05 to indicate optimism and anticipation.

Each sentence's affective profile is encoded as a normalized superposition:

$$|\,\psi_{emotion}\rangle = \alpha\,|\,positive\rangle + \beta\,|\,neutral\rangle + \gamma\,|\,negative\rangle,$$

where $|\,\alpha\,|^2 + |\,\beta\,|^2 + |\,\gamma\,|^2 = 1$.

4 Result

4.1 Overall Patterns: Multiple Framings and Frame Dynamics

The analysis confirmed that AI-generated news articles frequently exhibit "multiple framings" within their narratives. The multi-framing intensity averaged 0.7716 (on a 0–1 scale), suggesting that a single article typically encodes several possible interpretations simultaneously rather than offering a univocal story. This means that readers could plausibly reach different conclusions about events depending on which parts of the narrative they emphasize. Such a finding offers empirical support to the concept of quantum semantics in media: meanings remain in superposition until "collapsed" by reader interpretation. Rather than committing to one framing, AI-generated texts often include both optimism and skepticism, or conflict and harmony, side by side. This pattern challenges traditional expectations of objectivity in journalism and resonates with postmodern views of news as narrative construction, amplified by AI's probabilistic generation methods.

The comparative analysis between AI-generated content and CNA journalism reveals distinct differences in narrative structure and informational richness. In terms of Frame Competition, AI exhibits *perfect competition* (1.0000), meaning that all semantic frames coexist equally without dominance, reflecting a balanced multi-perspective discourse. In contrast, CNA demonstrates a *high but not perfect competition* (0.9173–0.9985), suggesting a slight frame hierarchy that produces more structured and coherent narratives. Examining von Neumann Entropy, AI maintains a consistent entropy of *4.0000*, indicating uniform information density and even distribution of meanings. CNA, however, shows *variable entropy* ranging from *3.4378 to 7.3508*, which is approximately 84% higher than AI, evidencing greater informational diversity and narrative complexity. Overall, CNA content is significantly more information-dense, while AI maintains ideal frame equality and supports multiple simultaneous interpretations. Both sources sustain a neutral tone, but CNA achieves neutrality through editorial consistency, whereas AI achieves it through semantic averaging. These findings suggest that AI-generated content successfully models the "multiple framings" phenomenon characteristic of pluralistic discourse, yet this comes at the cost of reduced information density compared to the more hierarchically structured and detail-rich style of professional journalism.

Frame analysis revealed an additional pattern: very high frame competition (average 0.8891) but low frame conflict (average 0.1640). AI news tends to present numerous frames simultaneously but arranges them to minimize contradiction. For instance, a controversial policy article might include both "public safety" and "personal freedom" frames without resolving which is correct. Each frame is presented discretely, often by different speakers, allowing peaceful coexistence. Unlike traditional journalism, where competing frames often clash, AI-generated narratives appear to place frames side by side. This reflects a distinctive "high competition, low conflict" framing style that broadens interpretive possibilities without forcing resolution.

From a persuasion standpoint, this polyvalence can be read as strategic ambiguity. By leaving interpretation open, AI news accommodates varied audiences, each of whom may find their own perspective validated.

4.2 Emotional Tone and Sentiment Use

Emotional tone analysis showed that AI-generated news maintains a largely neutral to slightly positive register, with negative sentiment being rare. The mean positive sentiment intensity was ~0.2065, while only about 23.4% of articles carried significant negative language. Neutrality dominated across the corpus, suggesting a style that favors factual exposition peppered with subtle positivity.

Breaking down by section revealed important differences. Titles carried the strongest emotional charge (average 0.2760), often employing dramatic or evaluative words such as "重大突破" (major breakthrough) or "嚴重警告" (stern warning). About 37% of titles included stronger sentiment than the body, aligning with journalistic practices of crafting attention-grabbing headlines.

Dialogues, which made up the body text, were the most neutral (average sentiment 0.1566). Emotive expressions were frequently balanced by counterpoints in simulated multi-speaker exchanges. This dynamic reduced variance and created an impression of neutrality, reinforcing credibility through balanced voices.

Descriptions were mostly factual, resembling wire-service summaries. When sentiment appeared, it leaned positive, often framing problems alongside hopeful solutions. For example, disaster coverage frequently pivoted to recovery measures, mitigating negativity.

Persuasively, this pattern suggests AI news seeks credibility through neutrality while using selective positivity to foster reassurance. Rather than overtly directing audience emotions, it subtly steers interpretation toward optimism.

4.3 Agenda Breadth and Information Density

Another key finding was the broad agenda breadth of AI-generated news. Articles often included wide-ranging contextual information but lacked strong emphasis on priority issues. Von Neumann entropy was highest in descriptions (0.8937), indicating dense, information-rich content. Titles, by contrast, had low entropy, while dialogues fell in between.

Descriptions also scored highest in frame competition (~0.9050). They frequently included multiple angles—political, economic, social, and historical—in a single paragraph. For example, a corporate scandal description referenced ethical implications, financial effects, prior incidents, and investor reactions, leaving the reader to decide which angle mattered most.

This encyclopedic style contrasts with traditional journalism, where editors foreground particular aspects to guide audience focus. AI-generated news instead outsources agenda-setting to readers by presenting numerous perspectives without hierarchy. From a persuasive standpoint, agenda breadth can increase credibility by conveying thoroughness but risks diluting focus. It may also create an "illusion of depth," where sheer quantity of details fosters trust even if no clear conclusion is provided.

5 Analysis

5.1 Rethinking Media Theory in a Quantum Framework

Classical theories of agenda-setting and framing assume linear effects: media highlight issues to shape public focus and frame them in ways that guide interpretation. AI-generated news disrupts this model. Instead of a singular agenda, AI texts exhibit agenda multiplicity—a wide array of issues included without a clear hierarchy. This suggests a need for an "algorithmic agenda-setting" concept, where priorities emerge from data frequency or algorithmic design rather than editorial judgment. Readers may be told "many things to think about" without guidance on which matter most.

Similarly, framing becomes pluralistic. Rather than privileging one interpretive angle, AI news embeds multiple frames within a single article. This polysemy resonates with postmodern media theory, particularly John Fiske's work on polysemic texts and Leah Ceccarelli's notion of strategic ambiguity. The AI is not a rhetor with intent, but the effect mirrors deliberate ambiguity: conflicting audiences can each find validation. A conservative and a liberal might interpret the same AI-generated political story differently, confirming their predispositions. This parallels theories of selective perception and hostile media effect, where ambiguity fosters divergent interpretations.

From a quantum perspective, meaning exists in superposition until "collapsed" by the reader. Different audiences measure the text differently, producing varied interpretations. Unlike traditional journalism, which assumes a preferred reading, AI-generated journalism may lack any singular intended meaning.

5.2 Strategic Ambiguity and Persuasion

Strategic ambiguity emerges as a core persuasive element. By presenting multiple perspectives without adjudication, AI-generated news broadens acceptability. Ceccarelli (1998) noted that ambiguity unites conflicting audiences, and AI articles function similarly. This ambiguity can diffuse polarization by avoiding outright conflict, but it also dilutes clarity. Readers may leave with less certainty about what truly matters.

The persuasive outcome is paradoxical: ambiguity may reduce backlash but also reduce impact. Articles that hedge on every angle may keep audiences engaged without deeply swaying them. In polarized environments, such ambiguity could stabilize discourse by avoiding provocation, but it may equally risk fostering complacency.

5.3 Emotional Tone and Comfort Bias

Emotional analysis revealed that AI-generated news leans heavily neutral to slightly positive, with negative sentiment rare. This positivity bias, though subtle, may enhance persuasion by creating psychological comfort. Readers often prefer constructive or optimistic narratives, making them more receptive. By emphasizing reforms or solutions, AI-generated articles may foster goodwill toward institutions and authorities.

At the same time, the absence of strong negative framing reduces the risk of outrage-based virality. This could make AI-generated news less prone to fueling polarization but also less effective at holding power accountable. In terms of ethics, neutrality and optimism may seem impartial, yet they introduce a subtle pro-status-quo bias.

5.4 Practical Implications: Media Literacy and Regulation

For media literacy, these findings imply that readers must learn to navigate ambiguity. Rather than identifying a single bias, they should detect multiple frames and question what is absent. Educators may teach critical reading strategies for AI-generated texts: What perspectives are included? Which are omitted? Who benefits from this framing?

For regulation, quantum semantic metrics could aid content moderation. High multi-framing scores might flag overly contradictory or confusing texts, while high competition paired with high conflict could indicate propagandistic extremes. Automated monitoring could complement fact-checking in identifying problematic AI-generated content at scale.

5.5 Comparing Quantum and Classical NLP Approaches

Classical NLP techniques like BERT and LDA are effective at identifying dominant topics and frames. They assign fixed labels or topic proportions to text, capturing surface-level patterns of content. However, they struggle to represent ambiguity (Liu et al., 2023), competing interpretations (Waldon et al., 2025), or the contextual dynamics of persuasion (Saha et al., 2021; Bozdag et al., 2025). When faced with contradictory signals, such as praise and criticism in the same sentence, classical models tend to average or disambiguate, forcing a singular reading.

QNLP, in contrast, encodes language as quantum states capable of representing multiple meanings simultaneously. Using superposition, QNLP captures coexisting frames; entanglement models dependencies between semantic elements; and measurement simulates reader-driven interpretation, collapsing the state to a context-specific meaning. These features enable QNLP to reflect the uncertainty and multiplicity inherent in persuasive language. Rather than replacing classical methods, QNLP complements them - adding depth in cases of ambiguity, strategic framing, or interpretive variability where classical NLP falls short.

6 Conclusion

Future research should investigate how audiences actually interpret multi-frame AI-generated news. Do readers experience it as balanced and informative, or as vague and non-committal? Controlled experiments could measure which frames readers recall, which interpretations they adopt, and whether strategic ambiguity unites audiences or simply enables selective perception. Comparative analyses with human-written news on identical events would also clarify systematic differences, such as AI's tendency toward broader context or more neutral tone. Expanding to larger, cross-lingual corpora across domains like finance, sports, and health would further test the generality of these patterns and identify whether cultural or stylistic contexts alter persuasive dynamics.

On the technical side, QNLP methods can be refined to enable automatic frame detection, with advances in quantum machine learning and hardware allowing the encoding and analysis of larger, more complex semantic states. Practical applications may include monitoring tools that use metrics such as multi-framing intensity and frame entanglement to flag overly ambiguous or potentially polarizing articles, assisting editors in enhancing clarity. Finally, the study underscores the need for ethical guidelines in AI journalism, ensuring that neutrality does not come at the expense of omitting critical moral or evaluative frames. In sensitive domains like public health, balancing neutrality with clarity is essential for trustworthy communication.

References

Agostino, C. J., Thien, Q. L., Apsel, M., Pak, D., Lesyk, E., & Majumdar, A. (2025). A quantum

semantic framework for natural language processing. arXiv preprint arXiv:2506.10077.
Al Giffari, H. A., & Dermawan, A. (2025). AI vs. Human-Led Da'wah: A Comparative Rhetorical Analysis of Islamic Preaching in the Digital Age. DINIKA: Academic Journal of Islamic Studies, 10(1), 107-130.
Bozdag, N. B., Mehri, S., Yang, X., Ha, H., Cheng, Z., Durmus, E., You, J., Ji, H., Tür, G., & Hakkani-Tür, D. (2025, April 23). Must Read: A systematic survey of computational persuasion (arXiv:2505.07775v1). arXiv. https://arxiv.org/abs/2505.07775v1
Campbell, C. H. (2023). Automated Journalism at the Intersection of Politics and Black Culture: The Battle Against Digital Hegemony. Lexington Books.
Carlson, M. (2018). The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. In Journalism in an era of big data (pp. 108-123). Routledge.
Ceccarelli, L. (1998). Polysemy: Multiple meanings in rhetorical criticism. Quarterly Journal of Speech, 84(4), 395-415.
Coecke, B., Sadrzadeh, M., & Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. arXiv preprint arXiv:1003.4394.
de-Lima-Santos, M. F., & Jamil, S. (2024). Bridging the AI divide: human and responsible AI in news and media industries. Emerging Media, 2(3), 335-346.
Diakopoulos, N. (2019). Automating the news: How algorithms are rewriting the media. Harvard University Press.
Du Bois, J. W. (2003). Discourse and grammar. In *The new psychology of language* (pp. 47-88). Psychology Press.
Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of Communication, 43(4), 51–58.
Esteban, A. J. (2024, December). From Rules to Meaning Making: Teaching Grammar through Discourse Analysis as an Approach. In Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation (pp. 1047-1054).
Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda?. PNAS nexus, 3(2).
Graefe, A. (2016). Guide to Automated Journalism. Tow Center for Digital Journalism.
Ioscote, F., Gonçalves, A., & Quadros, C. (2024). Artificial intelligence in journalism: A ten-year retrospective of scientific articles (2014–2023). Journalism and Media, 5(3), 873-891.
Kuku, D., Charlotte Ojukwu, N. N., Onyoko Omali, T., & Shestakova, A. (2025). Algorithmic Bias in AI-Driven News Production and Dissemination: The Dynamics of X Misinformation During Elections.
Lippmann, W. (2017). Public opinion. Routledge.
Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., ... & Choi, Y. (2023). We're afraid language models aren't modeling ambiguity. arXiv preprint arXiv:2304.14399.
Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. Scientific Reports, 14(1), 4692.
McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. Public opinion quarterly, 36(2), 176-187.
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.
Meichanetzidis, K., Gogioso, S., De Felice, G., Chiappori, N., Toumi, A., & Coecke, B. (2020). Quantum natural language processing on near-term quantum computers. arXiv preprint arXiv:2005.04147.
Nah, S., Luo, J., Kim, S., Chen, M., Mitson, R., & Joo, J. (2024). Algorithmic Bias or Algorithmic Reconstruction? A Comparative Analysis Between AI News and Human News. International Journal of Communication, 18.
Paviour, B. (2025, October 16). AI-generated news sites spout viral slop from forgotten URLs. Nieman Journalism Lab. Retrieved October 18, 2025, from https://www.niemanlab.org/2025/10/ai-generated-news-sites-spout-viral-slop-from-forgotten-urls/
Pazzaglia, S., Vendetti, V., Comencini, L. D., Deriu, F., & Modugno, V. (2025). Passing the Turing Test in Political Discourse: Fine-Tuning LLMs to Mimic Polarized Social Media Comments. arXiv preprint ArXiv:2506.14645 v1 [cs. CL], 17.
Reubold, J. (2023). The Democratization of News-Analysis and Behavior Modeling of Users in the Context of Online News Consumption (Doctoral dissertation, Dissertation, Karlsruhe, Karlsruher Institut für Technologie (KIT), 2022).
Saha, S., Kalra, K., Patwardhan, M., & Karande, S. (2021, December). Performance of BERT on Persuasion for Good. In Proceedings of the 18th International Conference on Natural Language Processing (ICON) (pp. 313-323).
Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. Nature Human Behaviour, 1-9.
Singh, K., & Ngu, W. (2025, May). Bias-Aware Agent: Enhancing Fairness in AI-Driven Knowledge Retrieval. In Companion Proceedings of the ACM on Web Conference 2025 (pp. 1705-1712).
Sui, M. (2025). Writing with emotion? Assessing emotional valence and appeals in AI-generated vs. human-written articles. AI & SOCIETY, 1-15.
The Newsreel Project Consortium. (2021). Newsreel2: New teaching fields for the next generation of journalists [Research report]. Erich Brost Institute

for International Journalism. https://doi.org/10.17877/DE290R-22455
van Dijk, T. A. (2015). Critical discourse analysis. The handbook of discourse analysis, 466-485.
Waldon, B., Schneider, N., Wilcox, E., Zeldes, A., & Tobia, K. (2025, February 3). Large language models for legal interpretation? Don't take their word for it. Georgetown Law Journal, 114 (forthcoming). https://doi.org/10.2139/ssrn.5123124
Wazni, H., Lo, K. I., McPheat, L., & Sadrzadeh, M. (2024). Large scale structure-aware pronoun resolution using quantum natural language processing. Quantum Machine Intelligence, 6(2), 60.
Widdows, D., Aboumrad, W., Kim, D., Ray, S., & Mei, J. (2024). Natural Language, AI, and Quantum Computing in 2024: Research Ingredients and Directions in QNLP. arXiv Prepr. arXiv2403, 19758.
Yin, R., & Liu, X. (2025, May). From Technological Alienation to Value Regression: Ethical Regulation of Algorithmic Bias in the Post-Truth Era. In 2025 IEEE 10th International Conference on Smart Cloud (SmartCloud) (pp. 32-37). IEEE.
Zhang, J., He, R., Guo, F., & Liu, C. (2024, March). Quantum interference model for semantic biases of glosses in word sense disambiguation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19551-19559).

Interpretation of the level of ANGER in discussion forum

Suet Ching Soon
National United University
ssoon@gm.nuu.edu.tw

摘要

在這個數位時代，人們可以輕易地接觸到眾多社群媒體平台，以快速交流或互動。本研究觀察表憤怒情緒「怒」在中文線上討論區中網路使用者的情緒表達。本研究語料取自臺灣的電子佈告欄系統（Bulletin Board System, BBS），該平台具對話性，但沒有表情符號可以直接傳達對話雙方的情緒。本研究從該平台擷取了 7,464 筆包含「怒」的語料，刪除不符的語料後，有效語料為 7,285 筆。本研究檢視語料中「怒」的語義與搭配詞分布。結果顯示，有超過四分之一高頻搭配詞的語料屬於「怒」的非常規用法，意即這些語料中的用法並不必然表達「憤怒」的情緒。我們從這些非常規用法的搭配詞中，可看到語義從情緒表達轉向為表行動積極性，甚至是動作程度強弱。分析結果可見「怒」不僅是情緒表達，也可延伸作為說服聽者同理說話者行為的標記。所有中文標題，中文摘要和中文關鍵字。

Abstract

In this digital era, people have easy access to a vast array of social media platforms for quick communication or interaction. The ways how online users conveyed their emotional expression attracted our interest. This present paper investigates the literal emotional expression of ANGER in Chinese online discussion forum, targeting the term *nu4* ‘angry/anger’. We referred to a Bulletin Board System (BBS) in Taiwan which is a conversation-like platform with no emoji icon to convey emotion directly. A collection of 7,464 instances were retrieved from the platform. After deducting noisy data, we looked into the meanings and distribution of *nu4* of the 7,285 instances. The analysis showed a significant number of instances (nearly a quarter of the high frequency instances) belong to the unconventional category where *nu4* does not necessarily refer to emotion since the meaning of anger has degraded. This finding prompted a further collocate analysis to examine the functional shift. In conclusions, from the collocates of these unconventional uses of *nu4*, it showed a semantic shift from emotional expression to pragmatic functions, signaling aggressiveness or indicating the intensity of an action, and carrying a persuasive function in online forums to demonstrate the shift of the speaker’s actions.

Keywords: ANGER, emotional expression, Chinese *NU4*

1 Introduction

Emotion is an abstract concept by nature which generally expressed in metaphorical forms and has received attentions in interpreting the emotion conceptualization (Kövesces 2005, Matsuki 1995, Lakoff and Kövesces 1987). It is believed that emotions are affected by cultures. For example, ANGER IS A HOT FLUID IN A CONTAINER or THE ANGRY PERSON IS A PRESSURED CONTAINER is a common metaphorical expression in English. Though many in the past (Yu, 1998; Chen, 2010, etc.) studied Chinese ANGER metaphors and found some commonly shared metaphors such as ANGER IS HEAT with other languages, source domains might slightly differ. Yu (1998) pointed out that instead of referring ANGER as HOT FLUID, Chinese tend to refer it as HOT GAS, i.e. reflecting the unique notion of “qi” in Chinese. Unlike these studies, our goal is to find out the representations of the emotion of ANGER in online discussion forum.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 369–374
November 20-22, 2025 ©2025 Association for Computational Linguistics

Emotions are part of the mental states. Croft (1993:64) said then following about mental state:

> There are two processes involved in possessing a mental state (and changing a mental state): (1) the experiencer must direct his or her attention to the stimulus, and (2) the stimulus (or some property of it) causes the experiencer to be (or enter into) a certain mental state.

For (1) in the except above, we get examples such as *He is angry at the comment*; and for (2) we get *The comment infuriates him* (our own examples). The two are respectively termed as 'subject-experiencer' and 'object-experiencer' (cf. Lakoff, 1971; Verhoeven, 2009) patterns. For Chinese, there are markers that indicate the direction of anger, namely *dui4* 'towards', *rang4* 'cause/make', *shi3* 'cause', etc. (see also Cheung and Larson, 2006). These markers, however, mainly work for conventional terms such as *sheng1qi4* and *fa1huo3* 'be.angry'. When it comes to our target term, *nu4* 'angry/anger', which is a direct expression of ANGER, we found some slightly different patterns.

Crucially, these linguistic realizations of anger in digital discourse are not only descriptive of emotional states but also potentially strategic in persuasion. Most studies of persuasion analyze how narratives are formalized, focusing on settings where a persuader utilizes a broad, sense-making explanation to frame their claim, influencing audience perception (O'Keefe, 2016; Bullock, Shulman, and Huskey, 2021; Hosman, 2002; Moyer-Gusé, 2008; Miller, 2002; Green and Brock, 2000, etc.). For instance, Schwartzstein and Sunderam (2021) provided evidence that persuasion increases when the messages lead audience to adopt a narrative supported by clearly presented and convincingly explained data. Conversely, Ispano (2022) suggested that audiences who focused strictly on a high coherence criterion in interpreting utterances may limit their understanding. O'Keefe (2016) argued that various forms of consequence-based arguments in persuasion research, though appearing quite different, share core principle that messages are more persuasive when outcomes are considered as desirable.

In contrast, evidence from online forums shows that *nu4* demonstrated a key pragmatic divergence that may shift away from denoting literal emotion toward signaling aggressiveness or intensity, thereby strengthening argumentative force. From a persuasive-language perspective in AI, such shifts are significant because they illustrate how emotional expressions function as rhetorical resources, enhancing stance-taking and influencing audience perception without relying solely on traditional, formal narratives, and often through nonverbal means (Mehrabian and Williams, 1969; Miller, 2002; Green and Brock, 2000, etc.). In fast-paced digital contexts, brief textual utterances, especially those that contain emotional expression, may prompt emotional engagement and draw the audiences in even without full story, and thus creating persuasive impact, even if it is in a subtle, non-traditional signals.

Consequently, for AI systems designed to process or generate persuasive language, understanding these nuanced uses of anger terms (whether literal or metaphorical, emotional or argumentative) becomes essential, as it shows how emotion expressions can be harnessed not only to convey states of mind but also to achieve persuasive impact in human-AI communication

2 *Nu4* in online discussion forum

Nu4 is an equivalent of 'anger' but it is used perhaps in a more serious manner. This corpus-based study not only look into the use and interpretation of *nu4* in online discussion forum, but also distinguishing the conventional use as an emotional ANGER expression from its unconventional applications. For the data of this study, a total of 7,464 instances of *nu4* were retrieved from the PTT corpus, a Bulletin Board System (BBS) in Taiwan (accessed through http://lopen.linguistics.ntu.edu.tw/copens/). The PTT was selected because it contains conversation-like discussion threads within which emotion was usually expressed directly. The communication style on this platform is more direct in a way that emotion is conveyed explicitly without relying those subtle cues or context. In this study, 179 instances were deleted which including movie title, proper nouns, or unidentifiable uses. After analysis, the results are shown below.

Categories	Sub-Categories	Sub-Categories Freq. (%)	Category Freq.	Total (%)
Lexical items	Resultative (*nu4* + verb) (including *chi1* 'to eat', *qiang4* 'to irritate' *pi1* 'to criticize', *he1* 'to drink', *shuai3* 'to fling', *chou1* 'to draw/pull out')	3491 (71.17%)	4905	65.72
	Angry (*be.angry/anger*)	1414 (28.83%)		
Emoticons			1571	21.04
Idiomatic Expression			810	10.84
Deleted instances (proper nouns or unidentifiable uses)			179	2.40
Total			**7464**	**100**

Table 1 Distribution of *NU4* in PTT

First, all the instances were categorized based on the appearance of *nu4* found in the data, including Idiomatic expressions and Lexical items both of which were defined by their semantic features, as well as Emoticons. Within these, sub-categories can be found under Lexical items, including Resultative and Angry (used either as verbs or nouns), which together accounted for over 65% of the total instances. In order to further understand the behavior of *nu4* used in online discussion forum, we further examined the collocates of *nu4*, particularly focusing on the category termed Resultative.

First of all, *nu4* is used 'lexically' in the instances, and we found that, still, a majority of the instances were used to express the meaning 'be.angry/anger'. As in example 1 below, the experiencer was angry because of the reason that s/he was not being informed or updated. Stimulus which causes the experiencer to be in the ANGER state could be observed in instances.

1. 老闆 還是 沒有 主動 告訴 我 任何
   boss still NEG. initiative to.tell 1SG any
   消息 我 真的 **怒**了
   news 1SG real ***NU4*-PTCL**
   The boss still hasn't informed me any news and I am really angry.

The use of *nu4* in this example denote the lexical emotion meaning 'be.angry'. The conventional uses of *nu4* refers to literal lexical expressions of emotion to convey the emotional states of 'be.angry/anger'. As for non-lexical or symbolic cues, such as text-based icons, emoticons, and punctuations would be categorized under Emoticon where emotion is conveyed visually through a range of symbolic strategies rather than lexical.

Yet, unlike the conventional use of 'to be angry', the use of *nu4* in this online discussion forum are often accompanied by a follow-up impulsive consequence (*nu4 shui4jiao4* 'angry-sleep', *nu4 chu2zhi2* 'angry-to.top.up.money'). It is worth noted that for these instances, they do not necessary highlight the emotional meaning of being angry or anger, but rather it aims to bring out the results or actions followed. These uses of *nu4* fall under Resultative category (constituted almost 46.76%) in this study.

2. 布丁 布丁 布丁 可是 我 在
   pudding pudding pudding but 1SG on
   減肥 嗚嗚嗚 算了
   lose- weight Wooocrying never-mind
   怒 睡覺
   NU4 sleep
   Pudding, pudding, pudding but I am on diet. (CRYING) Never mind, go to sleep.

3. 為了 湊到 16 隻 英雄 **怒**
   for-purpose to-collect 16 CL. hero ***NU4***
   儲值 值得 了
   top-up worth ASP.
   It is worthwhile topped-up in order to collect 16 heroes.

In example 2, *nu4* showed a shift in intention of the speaker, i.e. the experiencer failed to get what s/he wanted (i.e. the pudding) as s/he was on diet, s/he decided to go to sleep instead. However, as in example 3, *nu4* does not always refer to the emotion of the speaker or experiencer while the action was carrying out, but to emphasis the action of topping-up.[1] These examples show that the use of *nu4* in the online discussion forum change from denoting the emotion of being angry to referring to the intensity and aggressiveness of the action. We will discuss further the collocates of these unconventional uses of *nu4* in next section.

Meanwhile, we also found that *nu4* works almost similarly like an emoticon (21.04%), see 4 and 5.

4. 盯著 看 真的 是 服務 態度 有夠
   Stare-ASP see real is service attitude really
   差 (**怒**!!!!!!!!)
   bad (***NU4***)
   The service is really bad as they staring (at us) (EMOTICON$_{ANGER}$)

5. 多少 蚊子 啦!!!!!!!
   how-many mosquitoes-PTCL
   怒!!!!!! (╯ˋ□′)╯
   NU4
   How many mosquitoes? (EMOTION$_{ANGER}$)

For instances under this category, lexical *nu4* often appears accompanied by many exclamation marks, or in brackets, as in example 4. This serves as an expression to convey emotion in this text-based environment. Some would follow by text emoticon, as in example 5, to emphasize the intensity of emotion. Both will be categorized under Emoticon. As mentioned, there is no emoji icon to convey emotion on this platform, text-based emoticons are the creation of users. Thus, these emoticons picturized the abstract mental state of the experiencer, showing their emotion vividly. As for the remaining instances, mostly are idioms and fixed expressions (including those that were collected as part of the conceptual metaphors mentioned by others) contributed to nearly 11% of the total instances.

6. 當 悲劇 發生, 我們 悲
   when tragic happened 3PL. sadness
   不可抑, **怒** 不可止…
   unsuppress-able ***NU4*** un-cease-able
   When tragic happened, our sadness could not be suppressed, anger could not be ceased…

For idiomatic expression, *nu4* in this category keeps the literal definitions which refers to the conventional use of the emotion of anger. As in example 6, the experiencer could neither refrain their sadness nor could they cease their anger once the tragic happened. The conventional meaning of *nu4* in these idiomatic fixed expressions is highly predictable.

From the data retrieved, the conventional literal expression of *nu4* 'be.angry/anger' still makes up the majority. Meanwhile, a significant number of instances falling under Resultative category exhibited the unconventional use of *nu4*, which attracted our attention to look further into the collocates.

3 Unconventional use of *nu4*

For instances fall under Resultative category, collocate analysis was conducted to identify high frequency collocates of *nu4* in the discussion forum. According to the analysis of our data, the emotion of ANGER and the action of experiencer are closely related. It is interesting to point out that the emotion of ANGER declined in some of the instances. The high frequency collocates ($F \geqq 100$) comprised of verbs including *chi1* 'to eat' (612), *qiang4* 'to irritate' (490), *pi1* 'to criticize' (444), *he1* 'to drink' (112), *shuai3* 'to fling' (103), and *chou1* 'to draw/pull out' (103), constituted about 24.97% (1864) of all the instances.

Take *chi1* 'to eat' for example, which obtained the highest frequency in our data, the instances do not necessarily refer to 'angry-eat' neither to convey the emotion of ANGER as found in example 7, but to express the aggressiveness of the action taken, as in example 8 and 9.

7. 你 希望 可以 找 個 人 一起
   2SG. wish able to.find CL. people together
   瘋狂 的 唱歌、**怒** 吃 美食

[1] It is about an online-game which player might want to top-up credits in order to accomplish certain tasks within a short period of time. By doing so, gamers could also upgrade their weapons or equipment more quickly.

be.crazy DE to.sing ***NU4*** to.eat gourmet
四處走走
wandering.around
You wish you could find someone to sing wildly with, devour delicious food, and explore different places together.

8. 每 張 都 是 五 顆 石頭 要
each CL. ASP. SHI 5 CL. stone want-to
把 他們 **怒** 吃 還是 會 怕怕
BA. 3PL. ***NU4*** to.eat still AUX. scare
Each card worth 5 stone, it is still scary to eat them.

9. 剛 八點 左右 停電，
just 8 o'clock about power-failure
肥宅 我 沒 事 做， **怒**
fat-otaku 1SG. NEG. matter to.do ***NU4***
吃 兩 碗 泡麵
to.eat 2 CL. instance-noodle
There was a power failure at about 8 o'clock, as a homebody I have nothing to do so I ate 2 bowls of instance noodle.

For example 7, it implies the determined attitude and the extent degree of the speaker to enjoy gourmet food with the 'someone'. Also, analysis shows a total of 116 instances collocated with 'eat' were related to online-gaming, such as example 3 and 8. Gamers could 'combine' some of the available equipment in their platform lists in order to level-up their power/skills or to get access to next level. For gamers to obtain the crucial elements for level-up, one of the key methods is to collect as much the required items as possible in the game. And *chi1* 'to eat' is commonly used to refer to the action of collecting or combining treasure items in the game, so that the gamers could level up and unlock new abilities or new challenges. As in example10 and 11, instead of expressing the emotion of ANGER, *nu4* in these sentences tend to modify the assertive attitude of speaker.

10. 太 謝謝 了！！該
too thankful ASP. should
收心 了
back-to-work-mode ASP.
準備 **怒** 讀書 明天
prepare ***NU4*** study tomorrow
期中考 加油！！
mid-term make-effort
(I am) grateful. I should get back to work mode and to get ready study hard for tomorrow mid-term.

11. 10萬 夠 你 買 紅單
100 thousand enough 2SG. buy red-slip
轉手 **怒** 賺 30～200萬
resell ***NU4*** earn 300-2000 thousand
A hundred thousand is enough for you to buy a pre-sale house and resell it with 300-2000 thousand profit.

As in 10, the experiencer is suggesting themselves to get back to the 'back-to-work' mode where *nu4* du2shu 'angry-study' implies to study hard for the exam next day rather than 'to study angrily'. It indicated that the experiencer determined to study hard, as well as showing the assertive manner of the speaker.

In Example 11, the speaker commented that with a 100 thousand investments in buying the pre-sale housing, the audience, could easily make a profit of 300 thousand to 2 million. The use of *nu4* in this sentence modifies the action 'to-earn', implying that the earning profit of selling the pre-sale houses is considered relatively high from the speaker point of view. The function of *nu4* in this situation is to emphasize the extent or high level of the profit earned. This implies that the returns from this investment are significant and easily obtainable, sending a message that the investment is profitable and worth pursuing, and thus serving a persuasive function to encourage the audience to take action. This suggests that *nu4* is used not only to convey emotion, but also to express speaker's attitude and invite audience understanding regarding their proposed actions.

4 Conclusion

From this study, we found that the emotion of anger in the online discussion forum is different from the conventional metaphors found in the past. The morphological constructions also differ from how they are normally used in texts. Among these unconventional uses of *nu4*, the emotion meaning degraded while meanings emphasizing the degree and manner of actions extended. It is believed that the meanings of *nu4* shift from the emotion of ANGER to the expression of aggressiveness, and to the extent level. Emotional utterances and emoticons serve to hint at and trigger readers' or receivers' emotions, and thereby enhancing engagement with the message.

The use of *nu4* in the context not only expresses the aggressiveness of the action, meanwhile, it also serves as a marker that the speaker, reacting to the unexpected events, has shifted their intention toward carrying out the action. This, therefore, implies a subtle persuasive attitude in the context. It is worth looking into the expressions of extent level in Chinese in future studies.

Based on these findings, this study explores how emotional expressions evolve in online discussion forum particularly within text-based environment that lack of nonverbal cues. By examining the lexical and pragmatical shift of *nu4*, it gives clearer picture of how online users creatively adapt their emotional expression to communicate effectively.

Acknowledgments

We would like to thank the reviewers for their constructive feedback and thoughtful comments, which has been invaluable in improving this manuscript.

References

Albert Mehrabian and Martin Williams. 1969. Nonverbal concomitants of perceived and intended persuasiveness. *Journal of Personality and Social Psychology,* 13(1):37–58. https://doi.org/10.1037/h0027993

Alessandro Ispano. 2022. The perils of a coherent narrative. *THEMA Working Papers 2022-13, THEMA* (Théorie Economique, Modélisation et Applications), CY Cergy Paris Université, France.

Candice Cheung and Richard Larson. 2006. Chinese Psych Verbs and Covert Clausal Complementation. Paper Presented at *Chicago Workshop on Chinese LX.*

Daniel O'Keefe. 2013. The Relative Persuasiveness of Different Forms of Arguments-From-Consequences: A Review and Integration. Annals of the *International Communication Association*, 36(1):109-135. https://doi.org/10.1080/23808985.2013.11679128

Elisabeth Verhoeven. 2009. Subjects, agents, experiencers, and animates in competition: Modern Greek argument order. *Linguistische Berichte*. 219:355-376.

Emily Moyer-Gusé. 2008. Toward a Theory of Entertainment Persuasion: Explaining the Persuasive Effects of Entertainment-Education Messages. *Communication Theory*, 18:407-425.

George Lakoff. 1971. *Cross-over Phenomena*. New York: Holt, Rinehart and Winston.

George Lakoff and Zoltán Kövecses. 1987. The cognitive model of anger inherent in American English. In Dorothy Holland and Naomi Quinn (eds.). *Cultural Models in Language and Thought*. Cambridge: Cambridge University Press. pages 195-221.

Gerald R. Miller. 2002. On Being Persuaded: Some Basic Distinctions. In James Price Dillard and Michael Pfau (eds.), *The Persuasion Handbook: Developments in Theory and Practice*. London & New Delhi: Sage Publication, pages 3-6.

Joshua Schwartzstein and Adi Sunderam. 2021. Using Models to Persuade. *American Economic Review,* 111(1):276-323.

Keiko Matsuki. 1995. Metaphors of anger in Japanese. In: John R. Taylor and Robert E. MacLaury (eds.). *Language and the Cognitive Construal of the World*. Berlin, New York: De Gruyter Mouton, pages 137-152.

Lawrence A. Hosman. 2002. Language and Persuasion. In James Price Dillard and Michael Pfau (eds.), *The Persuasion Handbook: Developments in Theory and Practice*. London & New Delhi: Sage Publication, pages 371-389.

Melanie Green and Timothy Brock. 2000. The role of Transportation in the Persuasiveness of Public Narratives. *Journal of Personality and social Psychology*, 79(5):701-721.

Ning Yu. 1998. *The Contemporary Theory of Metaphor: A Perspective from Chinese*. Amsterdam/Philadelphia: John Benjamin's Publishing Company.

Olivia M. Bullock, Hillary C. Shulman, and Richard Huskey. 2021. Narratives are Persuasive Because They are Easier to Understand: Examining Processing Fluency as a Mechanism of Narrative Persuasion. *Frontiers in Communication*. 6:719615. https://doi.org/10.3389/fcomm.2021.719615

Peilei Chen. 2010. A Cognitive Study of "Anger" Metaphors in English and Chinese Idioms. *Asian Social Science*, 6(8):73-76.

William Croft. 1993. Case Marking and the Semantics of Mental Verbs. In James Pustejovsky (eds). *Semantics and The Lexicon*. Kluwer Academic Publishers, pages 55-72.

Zoltán Kövecses. 2000. The Concept of Anger: Universal or Cultural Specific? *Psychopathology*, 33:159-170.

ROCLING-2025 Shared Task: Chinese Dimensional Sentiment Analysis for Medical Self-Reflection Texts

Lung-Hao Lee[1,*], Tzu-Mi Lin[1], Hsiu-Min Shih[2],
Kuo-Kai Shyu[2], Anna S. Hsu[3], and Peih-Ying Lu[3]
[1]Institute of Artificial Intelligence Innovation, National Yang Ming Chiao Tung University
[2]Department of Electrical Engineering, National Central University
[3]Department of Medical Humanities and Education, Kaohsiung Medical University
*lhlee@nycu.edu.tw

Abstract

This paper describes the ROCLING-2025 shared task aimed at Chinese dimensional sentiment analysis for medical self-refection texts, including task organization, data preparation, performance metrics, and evaluation results. A total of six participating teams submitted results for techniques developed for valence-arousal intensity prediction. All datasets with gold standards and evaluation scripts used in this shared task are publicly available online for further research.

Keywords: dimensional sentiment analysis, valence-arousal intensity prediction, medical education, domain adaption, Chinese language processing

1 Introduction

In dimensional sentiment analysis, affective states are generally represented as continuous numerical values on multiple dimensions, such as valence-arousal (VA) space, as shown in Fig. 1 (Yu et al., 2016b). Based on this two-dimensional representation, any affective state can be represented as a point in the VA coordinate plane by determining the degrees of valence and arousal of given texts.

The existing methods for sentiment valence-arousal intensity prediction at different granularities from the word, phrase to text levels can be categorized as lexicon-based (Taboada et al., 2011; Thelwall et al., 2012; Paltoglou and Thelwall, 2013), regression-based (Wei et al., 2011; Malandrakis et al., 2013; Wang et al., 2016), neural-network-based (Kulshreshtha et al., 2018; Yu et al., 2018; Zhu et al., 2019; Yu et al., 2020; Deng et al., 2022), or transformer-based (Hung et al., 2021; Mukherjee et al., 2021; Park et al., 2021; Deng et al., 2023; Lin et al., 2023; Mendes and Martins, 2023). Recently, large language models (Liu et al., 2024; Xu et al., 2025) have also been used for sentiment intensity prediction with promising results.

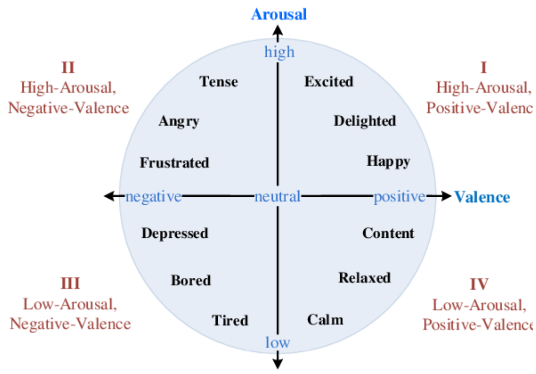


Figure 1: Two-dimensional valence-arousal space

The first dimensional sentiment analysis (DSA) task for Chinese words (Yu et al., 2016a) was organized at the IALP-2016 conference. The second edition was organized at the IJCNLP-2017 conference and included both Chinese words and phrases (Yu et al., 2017). The third edition was organized at the ROCLING-2021 conference to explore the sentence-level educational texts from students' self-evaluated comments (Yu et al., 2021). This year, we organized the fourth edition of the DSA task to analyze medical multi-sentence texts to describe doctors' self-reflection feelings.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 375–380
November 20-22, 2025 ©2025 Association for Computational Linguistics

Examples	Input & Output
Example 1	*Input*: ex01, 主治醫師曾經多次強調血液透析和輸血，以病人的狀況就是不建議，已經在加護病房積極治療了兩個禮拜，家屬卻遲遲無法達到共識。 (The attending physician has repeatedly emphasized that, given the patient’s condition, he/she does not recommend hemodialysis or blood transfusion. The patient has already been receiving intensive care in the ICU for two weeks, yet the family has been unable to reach a consensus.) *Output*: ex01, 4.750, 2.750
Example 2	*Input*: ex02, 視病如親，這個成語一直是一個難以達成的理想，但在 ICU 我感受到醫療端與病人和家屬站在同一陣線、共同努力對抗病魔，完成病人的願望的努力，讓我十分的動容。 (The saying ‘treat patients as if they were your own family’ has long been an admirable yet challenging ideal to realize. However, during my time in the ICU, I was deeply moved by the dedication of the medical team, who stood in solidarity with the patient and their family, working tirelessly together to combat illness and fulfill the patient’s final wishes.) *Output*: ex02, 6.900, 5.600

Table 1: Examples of the DSA-MST task.

The rest of this article is organized as follows. Section 2 provides a description of the Chinese Dimensional Sentiment Analysis for Medical Self-reflection Texts (DAS-MST) shared task. Section 3 introduces the constructed data sets. Section 4 describes the evaluation metrics. Section 5 compares evaluation results from the various participating teams. Finally, Section 6 provides conclusion and proposes future research directions.

2 Task Description

The goal of the DSA-MST shared task is to develop and evaluate the performance of Chinese sentiment analysis systems for multi-sentence texts written by doctors. The input is a self-reflective text describing a doctor’s feelings and opinions regarding his/her medical internship in Intensive Care Unit (ICU) rotation. The system should predict the real-valued valence-arousal (VA) intensity ratings using a nine-degree scale. A value of 1 on the valence and arousal dimensions respectively denotes extremely high-negative and most-calm sentiment, while a 9 denotes extremely high-positive and most-excited sentiment, and 5 denotes a neutral-valence and medium-arousal sentiment.

Example instances are presented in Table 1. The input format is the instance ID followed by given texts and the output format is the same ID, followed by valence and arousal ratings. In Example 1, the valence intensity is slightly negative at 4.75 and the arousal sentiment tends to be calm at 2.75. Example 2 shows a positive sentiment of 6.9 and medium-arousal of 5.6.

3 Data Preparation

The training set for this DSA-MST shared task is the Chinese EmoBank (Lee et al., 2022), a dimensional sentiment resource annotated with real-valued scores for both valence and arousal dimensions. The valence represents the degree of positive and negative sentiment, and arousal represents the degree of calm and excitement. Both dimensions range from 1 (highly negative or calm) to 9 (highly positive or excited). The Chinese EmoBank features various levels of text granularity including two lexicons called Chinese valence-arousal words (CVAW with 5,512 single

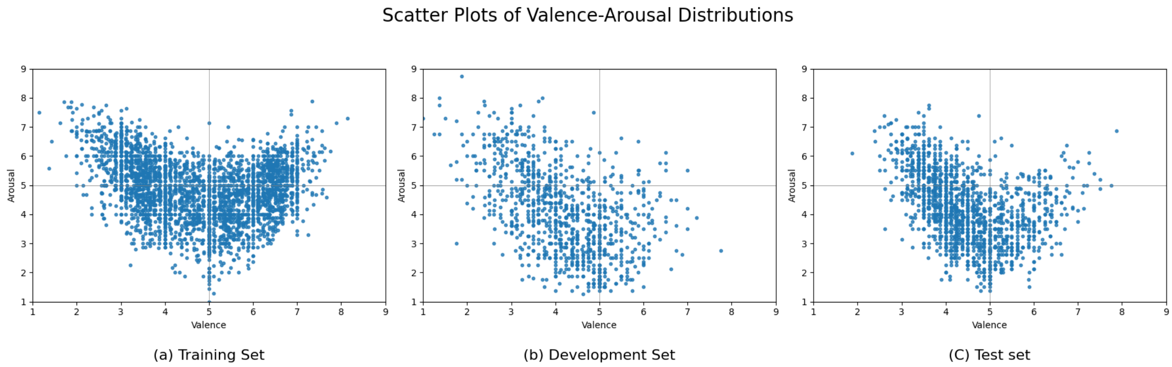


Figure 2: Scatter plots of valence-arousal distributions

words) and Chinese valence-arousal phrases (CVAP with 2,998 multi-word phrases), and two corpora called Chinese valence-arousal sentences (CVAS with 2,582 single sentences) and Chinese valence-arousal texts (CVAT with 2,969 multi-sentence texts).

The development and test sets consist of self-reflection texts written by doctors in their ICU rotation during their medical internship. The content covers the doctors' feelings and opinions towards patients and the patients' families. First, self-reflection texts were segmented into sentences and those containing sentiment words in the CVAW of Chinese EmoBank (Lee et al., 2022) were selected for manual annotation. Each sentence was presented to five Chinese native speakers for VA rating. Once the annotation process was finished, a cleanup procedure (Lee et al., 2022) was performed to remove outlier values which did not fall within 1.5 standard deviations (SD) of the mean. These outliers were then excluded from calculating the average VA values for each instance.

The annotated instances were randomly included in two mutually exclusive datasets. The development set contains 994 self-reflection texts (average 76.51 tokens) with VA ratings for system development, while the remaining 1,541 instances (average 76.81 tokens) were retained in the test set for system performance evaluation.

Figure 2 shows scatter plots of valence-arousal distributions, where the CVAT was used as the training set. Although they presented similar results, participating systems were allowed to use other publicly available data for prediction model learning, but such training data must be specified in the final system description.

4 Performance Metrics

System performance is evaluated by examining the difference between machine-predicted ratings and human-annotated ratings based on evaluation metrics including Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC), defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i|$$

$$PCC = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{a_i - \mu_A}{\sigma_A}\right)\left(\frac{p_i - \mu_P}{\sigma_P}\right)$$

where $a_i \in A$ and $p_i \in P$ respectively denote the i-th actual value and predicted value, n is the number of test samples, and μ_A and σ_A respectively represent the mean value and the standard deviation of A, while μ_P and σ_P respectively represent the mean value and the standard deviation of P.

The actual and predicted real values range from 1 to 9, so MAE measures the error rate in a range where the lowest value is 0 and the highest value is 8. The PCC is a value between −1 and 1 that measures the linear correlation between the actual value and the predicated value. A lower MAE and a higher PCC indicate more accurate prediction performance.

Each metric for the valence and arousal dimensions is ranked independently. A model's overall ranking is computed based on the mean rank across the four metrics. The lower the mean rank, the better the system performance.

Team (Submission)	Evaluation Metric				Overall Rank
	V-MAE (rank)	**V-PCC** (rank)	**A-MAE** (rank)	**A-PCC** (rank)	
CYUT-NLP (#356721)	**0.46** (1)	0.78 (2)	**0.74** (1)	**0.63** (1)	1
TCU (#356930)	**0.46** (1)	**0.81** (1)	0.76 (2)	0.61(2)	2
NTULAW (#357770)	0.50 (3)	0.75 (5)	0.79 (3)	0.59 (3)	3
SCUNLP (#357007)	0.51 (4)	0.76 (3)	0.87 (5)	0.59 (3)	4
KOLab (#358133)	0.53 (5)	0.76 (3)	0.82 (4)	0.58 (5)	5
HeyVergil (#356794)	0.63 (6)	0.62 (6)	1.01(6)	0.21 (6)	6

Table 2: Evaluation results of the DSA-MST task.

5 Evaluation Results

A total of six teams provided submissions to the leaderboard and submitted their technical papers. CYUT-NLP (Jian et al., 2025) applied the retrieval-augmented generation (RAG) and pseudo-labeling techniques to generate augmented data, and then used fine-tuned transformer-based models to predict VA ratings. TCU (Li and Lin, 2025) used several large language models (LLM) to extract contextual embedding representations and then fed semantic vectors into a regression model for VA rating prediction. The averaging ensemble technique was applied to assemble multiple prediction models for performance enhancement. NTULAW (Huang and Shao, 2025) fused encoders trained at different levels of granularity including word, phrase, and sentence to independently predict valence and arousal intensity. SCU-NLP (Pan and Wu, 2025) presented a dual-layer agent-executor framework for dimensional sentiment analysis. KOLab (Chan et al., 2025) and HeyVergil (Lin et al., 2025) systems were mainly based on BERT (Devlin et al., 2019) transformer fine-tuning for VA score prediction.

Table 2 shows the evaluation results. For the valence dimension, the best MAE of 0.46 and PCC of 0.81 was achieved by the TCU team (Li and Lin, 2025). For the arousal dimension, the best MAE of 0.74 and PCC of 0.63 was achieved by the CYUT-NLP system (Jian et al., 2025). In summary, the overall best results were provided by CYUT-NLP, followed by TCU and NTULAW (Huang and Shao, 2025).

6 Conclusion and Future Work

This paper provides an overview of the ROCLING-2025 shared DSA-MST task for Chinese dimensional sentiment analysis for medical self-reflection texts, including task descriptions, data preparation, performance metrics and evaluation results. Regardless of actual performance, all submissions contribute to the development of effective DSA systems in the medical domain, and each system description paper for this shared task also provides useful insights for further research.

We hope the data sets collected and annotated for this shared task can facilitate and expedite future development of Chinese DSA. The gold standard and evaluation scripts are made publicly available in a GitHub repository at https://github.com/NYCU-NLP/ROCLING-2025-ST-DSA-MST

Future directions will focus on the development of a Chinese domain-specific DSA. We plan to build new resources to develop techniques for the future enrichment of this research topic, especially for valence-arousal datasets in new domains.

Acknowledgments

This work was partially supported by the National Science and Technology Council, Taiwan under grant NSTC 111-2628-E-A49-029-MY3 and NSTC 114-2221-E-A49-059-MY3. This work was financially supported by the Co-creation Platform of the Industry-Ademia Innovation School, National Yang Ming Chiao Tung University.

References

Chia-Yu Chan, Chia-Wen Wang, and Jui-Feng Yeh. 2025. KOLab at ROCLING-2025 shared task: Research on emotional dimensions in Chinese medical self-reflection texts. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Yu-Chih Deng, Cheng-Yu Tsai, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2022. Predicting Chinese phrase-level sentiment intensity in valence-arousal dimensions with linguistic dependency features. *IEEE Access*, 10:126612-126620. https://doi.org/10.1109/ACCESS.2022.3226243

Yu-Chih Deng, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2023. Towards transformer fusions for Chinese sentiment intensity prediction in valence-arousal dimensions. *IEEE Access*, 11:109974-109982. https://doi.org/10.1109/ACCESS.2023.3322436

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171-4186. https://doi.org/10.18653/v1/N19-1423

Sieh-Chuen Huang, and Hsuan-Lei Shao. 2025. NTULAW at ROCLING-2025 shared task: Domain-adaptive modeling of implicit emotions in medical reflection. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Man-Chen Hung, Chao-Yi Chen, Pin-Jung Chen, and Lung-Hao Lee. 2021. NCU-NLP at ROCLING-2021 shared task: Using MacBERT transformers for dimensional sentiment analysis. *In Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing*, pages 380–384.

Yi-Min Jian, An Yu Hsiao and Shih-Hung Wu. 2025. CYUT-NLP at ROCLING-2025 shared task: Valence-arousal prediction in physicians' texts using BERT, RAG, and multi-teacher pseudo-labeling. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Devang Kulshreshtha, Pranav Goel, and Anil Kumar Singh. 2018. How emotional are you? Neural architectures for emotion intensity prediction in microblogs. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2914–2926.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions of Asian and Low-Resource Language Information Processing*, 21(4), article 65: 1-18. https://doi.org/10.1145/3489141

Hsin-Chieh Li, and Wen-Cheng Lin. 2025. TCU at ROCLING-2025 shared task: Leveraging LLM embeddings and ensemble regression for Chinese dimensional sentiment analysis. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Tzu-Mi Lin, Jung-Ying Chang, and Lung-Hao Lee. 2023. NCUEE-NLP at WASSA 2023 Empathy, Emotion, and Personality Shared Task: Perceived intensity prediction using sentiment-enhanced RoBERTa transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 548-552. https://doi.org/10.18653/v1/2023.wassa-1.49

Ting-Yi Lin, Cong-Ying Lin, and Jui-Feng Yeh. 2025. HeyVergil at ROCLING-2025 shared task: Emotion-space-based system for doctors' self-reflection sentiment analysis. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487-5496. https://doi.org/10.1145/3637528.3671552

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, Language Processing*, 21(11):2379–2392. https://doi.org/10.1109/TASL.2013.2277931

Goncalo Azevedo Mendes, and Bruno Martins. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *Proceedings of the 45th European Conference on Information Retrieval*. https://doi.org/10.1007/978-3-031-28244-7_6

Rajdeep Mukherjee, Atharva Naik, Sriyash Poddar, Soham Dasgupta, and Niloy. Ganguly. 2021. Understanding the role of affect dimensions in detecting emotions from tweets: A multi-task approach. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2303–2307. https://doi.org/10.1145/3404835.3463080

Georgios Paltoglou, and Michael Thelwall. 2013. Seeing stars of valence and arousal in blog posts. *IEEE Transactions on Affective Computing*, 4 (1): 116–123. https://doi.org/10.1109/T-AFFC.2012.36

Hong Rui Pan, and Jheng-Long Wu. 2025. SCUNLP at ROCLING-2025 shared task: Systematic guideline refinement for continuous value prediction with outlier-driven LLM feedback. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4367–4380. https://doi.org/10.18653/v1/2021.emnlp-main.358

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307. https://doi.org/10.1162/COLI_a_00049

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173. https://doi.org/10.1002/asi.21662

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Community-based weighted graph model for valence-arousal prediction of affective words. *IEEE/ACM Transactions on Audio, Speech, Language Processing*, 24(11): 1957–1968. https://doi.org/10.1109/TASLP.2016.2594287

Wen-Li Wei, Chung-Hsien Wu, and Jen-Chun Lin. 2011. A regression approach to affective rating of Chinese words from ANEW. In *Proceedings of the International Conference on Affective Computing and Intelligent Systems*, pages 121-131. https://doi.org/10.1007/978-3-642-24571-8_13

Zhe-Yu Xu, Yu-Hsin Wu, and Lung-Hao Lee. 2025. NYCU-NLP at SemEval-2025 Task 11: Assembling small language models for multilabel emotion detection and intensity prediction. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, pages 1129–1135.

Liang-Chih Yu, Lung-Hao Lee, and Kam-Fai Wong. 2016a. Overview of the IALP 2016 shared task on dimensional sentiment analysis for Chinese words. In *Proceedings of the 20th International Conference on Asian Language Processing*, pages 156-160.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016b. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540-545. https://doi.org/10.18653/v1/N16-1066

Liang-Chih Yu, Lung-Hao Lee, Jin Wang, and Kam-Fai Wong. 2017. IJCNLP-2017 Task 2: Dimensional sentiment analysis for Chinese phrases. In *Proceedings of the 8th International Joint Conference on Natural Language Processing: Shared Tasks*, pages 9-16.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2018. Refining word embeddings using intensity scores for sentiment analysis. *IEEE/ACM Transaction on Audio, Speech, Language Processing*, 26(3):671–681. https://doi.org/10.1109/TASLP.2017.2788182

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2020. Pipelined neural networks for phrase-level sentiment intensity prediction. *IEEE Transactions on Affective Computing*, 11(3): 447–458. https://doi.org/10.1109/TAFFC.2018.2807819

Liang-Chih Yu, Jin Wang, Bo Peng, Chu-Ren Huang. 2021. ROCLING-2021 shared task: Dimensional sentiment analysis for educational texts. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing*, pages 385-388.

Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial attention modeling for multi-dimensional emotion regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 471–480. https://doi.org/10.18653/v1/P19-1045

CYUT-NLP at ROCLING-2025 Shared Task: Valence–Arousal Prediction in Physicians' Texts Using BERT, RAG, and Multi-Teacher Pseudo-Labeling

JIAN, YI-MIN
Department of CSIE
Chaoyang University of Technology Taichung, Taiwan
s11427601@gm.cyut.edu.tw

HSIAO An Yu
Department of CSIE
Chaoyang University of Technology Taichung, Taiwan
s11227617@gm.cut.edu.tw

Shih-Hung Wu*
Department of CSIE
Chaoyang University of Technology Taichung, Taiwan
shwu@cyut.edu.tw

Abstract

Accurately modeling physicians' emotional states from self-reflection texts remains challenging due to the low-resource, domain-specific nature of medical corpora. The proposed workflow performs Retrieval-Augmented Generation (RAG) and multi-teacher pseudo-labeling to generate high-quality augmented data. This workflow enables effective cross-domain adaptation from general text corpora to professional medical texts. Evaluations on the ROCLING 2025 test set demonstrate improvements over the best-performing baseline in Valence–Arousal prediction accuracy and model stability. Importantly, the workflow is domain-agnostic and provides a generalizable methodology for systematically transferring models to new, low-resource domains, making it applicable beyond medical text analysis.

Keywords: RAG, BERT, pseudo-labeling

1 Introduction

In clinical healthcare settings, physicians are often exposed to high-pressure and high-risk working environments. Their emotional states not only influence the quality of clinical decision-making and patient care outcomes but are also closely related to their psychological well-being and professional development. Physicians' self-reflection texts provide an authentic record of their psychological experiences and emotional fluctuations. Automated analysis of these texts can facilitate emotional awareness among physicians, support hospital management decisions, and even enhance the quality of medical education and assessment.

Among various sentiment analysis approaches, traditional binary classification (e.g., positive/negative) or unidimensional scales are insufficient to capture the complex emotions commonly observed in clinical contexts, such as a "sense of heaviness in professional practice" or "perseverance amidst exhaustion." In contrast, the Valence–Arousal (V-A) two-dimensional model (Russell, 1980) can simultaneously measure the pleasantness and activation levels of emotions, providing a more nuanced representation of affective content.

Previous studies have confirmed the effectiveness of the V-A model in lexical and textual sentiment analysis (Wei et al., 2011; Wang et al., 2016a; Wu et al., 2017). In the field of Chinese sentiment analysis, Dimensional Sentiment Analysis (DSA) was first introduced at IALP 2016 (Wang et al., 2016b) and later extended to words and phrases at IJCNLP 2017 (Yu et al., 2017), followed by applications on student self-assessment texts ROCLING-2021 shared Task(Yu et al., 2021). Although these studies performed well on educational or general-domain corpora, models typically exhibit poor generalization in professional domains due to vocabulary differences, stylistic variations, and the complexity of domain-specific emotions.

The ROCLING 2025 shared task (Lee et al., 2025) applied DSA to physicians' self-reflection texts for the first time, requiring models to predict Valence and Arousal scores using Chinese EmoBank (Lee et al., 2022) as the primary data source. Compared with previous datasets, physicians' texts contain richer, multi-layered

* corresponding author

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 381–389
November 20-22, 2025 ©2025 Association for Computational Linguistics

emotions, such as uncertainty in clinical decision-making, professional responsibility, and emotional tension in doctor–patient interactions. This domain shift poses challenges to existing models and highlights the importance of cross-domain adaptation and low-resource learning strategies.

To address this challenge, this study proposes a three-stage data augmentation framework combining Retrieval-Augmented Generation (RAG, Lewis et al., 2020) and Multi-Teacher Pseudo-Labeling (Nguyen et al., 2024). In the first stage, large language models (LLMs) combined with RAG and few-shot learning generate medical texts consistent in style and context. In the second stage, BERT-based teacher models annotate the generated texts with V-A pseudo-labels. In the third stage, high-quality augmented datasets are constructed through consistency verification and outlier removal and are then used to train downstream models.

The core contribution of this study lies not only in generating high-quality augmented data but also in demonstrating how models can be effectively transferred from general corpora to professional medical domains, providing an empirical example of cross-domain adaptation in dimensional sentiment analysis. Experimental results show that models trained with augmented data perform comparably—or even better—on physicians' texts compared with models trained solely on original data, demonstrating successful domain adaptation. Additionally, the two high-quality augmented datasets produced in this study provide valuable resources for future research on Chinese medical text sentiment analysis and low-resource cross-domain applications.

The main contributions of this study are as follows:

We propose and implement a three-stage data augmentation framework combining RAG and Multi-Teacher Pseudo-Labeling, specifically designed for V-A sentiment analysis of physicians' self-reflection texts.

We demonstrate strategies and empirical results for effectively transferring models from general corpora to the professional medical domain, providing a reference for cross-domain adaptation.

We produce two high-quality augmented datasets that can serve as valuable resources for future Chinese medical text sentiment analysis and low-resource research.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 introduces the proposed methodology; Section 4 presents experimental design and results; Section 5 concludes; and Section 6 discusses future directions.

2 Research Background

2.1 Comparison Between ROCLING 2021 Shared Task and ROCLING 2025 Shared Task

The ROCLING 2021 shared task focused on students' self-evaluation texts, in which emotional expressions were mostly related to course progress, learning new knowledge, or difficulties in comprehension. These expressions were relatively straightforward in meaning. In contrast, the ROCLING 2025 shared task (Lee et al., 2025) is the first to apply dimensional sentiment analysis to doctors' self-reflection texts, which often convey more complex emotions, such as uncertainty in clinical decision-making, a strong sense of professional responsibility, and the emotional tension inherent in doctor–patient interactions.

Both tasks adopt the same Valence–Arousal annotation scheme (real-valued scores ranging from 1 to 9, with the same input/output format). However, the shift in domain substantially increases the level of difficulty. While students' texts generally express emotions more directly, making valence easier to identify, doctors' texts frequently exhibit multi-layered or mixed emotions, which makes arousal prediction considerably more challenging.

2.2 Dimensional Model of Emotion: Valence–Arousal

Emotion, as a complex psychological and social phenomenon, has long been a central topic in both psychology and natural language processing (NLP). Traditional emotion classification models, such as Ekman's six basic emotions—happiness, anger, sadness, pleasure, surprise, and fear (Ekman, 1992)—can categorize emotional states effectively, but they are limited in capturing multidimensional and continuous affective experiences.

To overcome these limitations, Russell (1980) proposed the Circumplex Model of Affect, which maps emotions onto a two-dimensional continuous space. The valence dimension reflects the pleasantness of an emotion, ranging from negative (unpleasant) to positive (pleasant), while the arousal dimension indicates the level of emotional activation or energy, ranging from low (calm) to high (excited). This model can capture nuanced emotional states, such as “calm joy” (high valence, low arousal) or “agitated anger” (low valence, high arousal), offering a more precise representation than traditional unidimensional sentiment classification for NLP tasks requiring fine-grained affective understanding (Schouten and Frasincar, 2015).

2.3 Retrieval-Augmented Generation (RAG) and Data Augmentation

In low-resource settings, data augmentation is a key strategy for improving model performance. Traditional techniques, such as synonym replacement (Wei and Zou, 2019), can expand the size of the training corpus but often suffer from contextual inconsistency or unnatural outputs, limiting their effectiveness in downstream tasks. The recent emergence of Large Language Models (LLMs) provides new opportunities for data augmentation, as these models can generate fluent and semantically diverse synthetic texts.

However, relying solely on LLMs may result in hallucinations, producing outputs that deviate from domain-specific contexts or contain factually incorrect information (Ji et al., 2023). To mitigate this problem, Retrieval-Augmented Generation (RAG) has been proposed (Lewis et al., 2020). RAG first retrieves relevant content from an external knowledge base or task-specific dataset and provides these retrieved examples as context to guide the LLM’s generation. By anchoring outputs to the target domain, RAG enhances contextual consistency, stylistic alignment, and factual accuracy, while preserving the linguistic diversity and fluency of LLM-generated text.

In this study, we adopt RAG using the DSAMST-Validation Set as the retrieval corpus, guiding the LLM to generate synthetic texts that more closely resemble the style and context of physicians’ self-reflections.

2.4 Pseudo-Labeling and Multi-Teacher Strategy

Pseudo-labeling is a widely used semi-supervised learning technique that leverages large amounts of unlabeled data to enhance model training. The typical procedure involves first training a teacher model on a small labeled dataset, then using it to predict labels for unlabeled data. High-confidence predictions are treated as pseudo-labels to expand the training set. While effective in increasing data utilization, relying on a single teacher model can introduce bias or errors, potentially misleading downstream student models.

To address this issue, a multi-teacher strategy is employed, which aggregates predictions from multiple teacher models to improve the robustness and reliability of pseudo-labeling (Nguyen et al., 2024). This approach often incorporates consistency checks and outlier removal, such as anomaly detection based on mean and standard deviation (Lee et al., 2022), to filter inconsistent or unreliable pseudo-labels. By applying these strategies, the quality of augmented datasets is enhanced, which in turn improves the generalization capability of downstream models.

3 Methods

3.1 Methodological Framework

This study focuses on predicting continuous valence and arousal values from physicians’ self-reflection texts and designs a strategy combining data augmentation and multi-teacher pseudo-labeling to enhance model generalization in low-resource scenarios. The overall methodology is divided into three main stages: data augmentation, teacher model training, and annotation and corpus cleanup. Ultimately, we construct two high-quality augmented datasets, which are then applied to downstream model training and performance comparison. We illustrate the overall workflow in Figure 1.

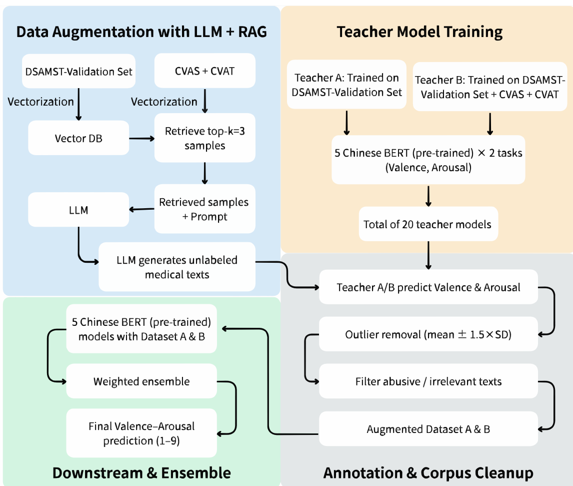


Figure 1. Methodological Framework for Data Augmentation, Teacher Annotation, and Ensemble-Based V-A Prediction.

3.2 Data Augmentation with LLM + RAG

In the data augmentation stage, we first vectorize the DSAMST-Validation Set provided by the ROCLING 2025 shared task and construct a vector database using the BAAI/bge-base-zh-v1.5 Chinese embedding model.

We retrieve the top three most similar samples from the database for each text in the Chinese EmoBank (Lee et al., 2022) derived from CVAS and CVAT and embed them as few-shot examples in the prompt to guide the large language model (LLM) in generating unlabeled medical texts that are consistent in style and context. The generated texts are guided by the following principles:

- Professionalism: Maintain domain-specific medical expression and avoid excessive colloquialism.
- Diversity: Introduce moderate variation in semantics and expression compared to retrieved examples to enhance corpus diversity.
- Authenticity: Avoid content that deviates from clinical context or is irrelevant to the task.

3.3 Teacher Model Training

We design two types of training datasets to annotate the augmented data for teacher models:

- Teacher A: DSAMST-Validation Set only.
- Teacher B: A combination of DSAMST-Validation Set, CVAS, and CVAT.

We fine-tune five pre-trained Chinese BERT models for each dataset:

1. bert-base-chinese
2. hfl/chinese-bert-wwm
3. hfl/chinese-roberta-wwm-ext
4. hfl/chinese-macbert-base
5. freedomking/mc-bert

We train two sub-models for each model and dataset because valence and arousal are two independent continuous prediction tasks. The final number of teacher models is as follows:

- Teacher A: 5 models × 2 tasks = 10 models
- Teacher B: 5 models × 2 tasks = 10 models
- Total: 20 teacher models

We fine-tune all models with an 80:20 training/validation split and use mean absolute error (MAE) as the loss function.

3.4 Pseudo-labeling and Corpus Cleanup

During annotation, we input the unlabeled augmented texts into both Teacher A and Teacher B models to obtain predicted valence and arousal values. To ensure reliability, we apply an outlier removal procedure similar to Chinese EmoBank (Lee et al., 2022):

Calculate the mean and standard deviation (SD) for each data point's predictions, remove outliers outside the range of mean ± 1.5 × SD, recalculate the mean of the remaining predictions as the final label.

Additionally, we remove any generated text that contains abusive, discriminatory, or clearly task-irrelevant content. Valence and arousal cleanup and calculations are performed independently to ensure annotation precision. After this process, two augmented datasets corresponding to Teacher A and Teacher B are generated for downstream model training and performance comparison.

3.5 Ensemble Strategy

To further improve predictive performance, this study employs a weighted ensemble approach to combine the outputs of multiple models. We first evaluate each model's performance on the validation set using mean absolute error (MAE) and Pearson Correlation Coefficient (PCC).

The initial weight of each model is defined as where W_i denotes the weight assigned to model i.

To ensure comparability across models, the weights are normalized, where M is the number of models in the ensemble and W_i' represents the normalized weight for model i.

Finally, the ensemble prediction is obtained via a weighted average, where $\hat{y}_i$ is the prediction of model i.

We integrate the predictions in this manner, effectively leveraging complementary information across individual models.

As a result, it enhances the accuracy and stability of the final predictions.

- Weighted Ensemble Initial Weight:

$$W_i = \frac{PCC_i}{MAE_i} \tag{1}$$

- Normalized Weight:

$$W_i' = \frac{W_i}{\sum_{j=1}^{M} W_j} \quad (2)$$

- Ensemble Prediction:

$$\hat{y}_{ensemble} = \sum_{i=1}^{M} W_i' \hat{y}_i \quad (3)$$

4 Experiments and Results

4.1 Training Data

This study utilizes three types of Chinese emotion datasets for model training and performance evaluation:

DSAMST-Validation Set: Derived from physicians' self-reflection texts, containing precisely annotated valence and arousal values (range 1–9). This dataset is primarily used for fine-tuning teacher models and serves as the basis for generating augmented data.

CVAS and CVAT (Lee et al., 2022): Contain valence and arousal annotations for single-sentence (CVAS) and short-text (CVAT) samples, respectively. These datasets are used to expand the training data for teacher models, enhancing their generalization to different text lengths and expression styles.

RAG-Augmented Dataset: Generated using the Retrieval-Augmented Generation (RAG) approach combined with few-shot LLMs, and annotated and cleaned by teacher models to form high-quality augmented data. The augmented datasets are categorized based on the source teacher models:

- Teacher A Augmented Data: Annotated by teacher models trained only on the DSAMST-Validation Set, focusing on the specific emotional distribution of physicians' self-reflection texts.

- Teacher B Augmented Data: Annotated by models trained on DSAMST-Validation Set combined with CVAS and CVAT, enhancing diversity and cross-style adaptability.

Dataset	Samples
DSAMST-Validation Set	994
Chinese Emobank-CVAS	2583
Chinese Emobank-CVAT	2926
RAG-CVAS-A Teacher	2583
RAG-CVAT-A Teacher	2926
RAG-CVAS-B Teacher	2583
RAG-CVAT-B Teacher	2926
ROCLING 2025 Test set	1541

Table 1. Dataset and Number of Samples

4.2 Evaluation Metrics

We assess model performance using Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC).

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |A_i - P_i| \quad (4)$$

- Pearson Correlation Coefficient (PCC):

$$PCC = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{A_i - \bar{A}}{\sigma_A}\right)\left(\frac{P_i - \bar{P}}{\sigma_P}\right) \quad (5)$$

where A_i and P_i denote the ground-truth and predicted values for sample i, and $\bar{A}$, $\bar{P}$, σ_A, σ_P are the corresponding means and standard deviations.

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.6550	0.6942	**1.0690**	0.4711
freedomking/mc-bert	**0.6093**	0.7180	1.0931	0.4843
hfl/chinese-bert-wwm	0.6289	0.6932	1.1268	0.4450
hfl/chinese-macbert-base	0.5979	**0.7237**	1.0761	0.4675
hfl/chinese-roberta-wwm-ext	0.6286	0.7048	1.0804	**0.5332**

Table 2. Performance of BERT models trained on CVAT+CVAS and evaluated on DSAMST-Validation Set

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.5182	0.7887	0.9499	**0.6258**
freedomking/mc-bert	0.5155	0.7839	0.9660	0.6173
hfl/chinese-bert-wwm	**0.5085**	**0.7909**	**0.9440**	0.6237
hfl/chinese-macbert-base	0.5154	0.7817	0.9533	0.6171
hfl/chinese-roberta-wwm-ext	0.5119	0.7898	0.9487	0.6178

Table 3. Performance of five BERT models trained on RAG-augmented data annotated by Teacher A, evaluated on DSAMST-Validation Set

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	**0.5099**	0.7843	0.9905	0.5967
freedomking/mc-bert	0.5213	0.7847	**0.9782**	0.5925
hfl/chinese-bert-wwm	0.5207	0.7812	0.9971	0.5949
hfl/chinese-macbert-base	0.5150	**0.7875**	0.9984	0.6016
hfl/chinese-roberta-wwm-ext	0.5205	0.7840	0.9850	**0.6021**

Table 4. Performance of five BERT models trained on RAG-augmented data annotated by Teacher B, evaluated on DSAMST-Validation Set

4.3 Cross-Domain Performance: Original vs. RAG-Augmented Data

From the results presented in Tables 2 to 4, we observe that when we train models on the original datasets (CVAT + CVAS) and directly evaluate them on physicians' self-reflection texts (DSAMST-Validation Set, Table 2), both Valence and Arousal MAE remain notably high, while PCC values are relatively low. This indicates that our models perform poorly under cross-domain conditions. In other words, training solely on general medical texts makes it difficult for the models to adequately capture the multi-layered and mixed emotional features present in physicians' texts.

In contrast, Tables 3 and 4 present the results of models we train on the RAG-augmented datasets proposed in this study, which combine LLM + RAG generation with multi-teacher pseudo-labeling. We observe that Valence MAE significantly decreases and PCC markedly improves, demonstrating that our approach enhances predictive performance for professional medical texts. Although the improvement in Arousal MAE is relatively modest, we find that the overall trend still surpasses the performance of models trained solely on the original datasets, indicating that our augmented data effectively provide samples similar in style and emotional distribution to the target domain.

Furthermore, when we compare Table 3 (Teacher A pseudo-labeled data) and Table 4 (Teacher B pseudo-labeled data), we observe that the models show similar performance on Valence prediction, while Teacher B achieves slightly better results on some Arousal metrics, reflecting the contribution of the multi-teacher strategy to the quality of augmented data.

Overall, these results demonstrate that our RAG-generated professional medical texts effectively mitigate the limitations of cross-domain data scarcity and substantially enhance model generalization in the professional medical domain, highlighting the innovative contribution of our study to cross-domain adaptation.

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.5	**0.77**	0.79	0.57
freedomking/mc-bert	0.5	**0.77**	**0.78**	**0.59**
hfl/chinese-bert-wwm	**0.48**	0.74	0.8	0.56
hfl/chinese-macbert-base	0.49	0.76	0.8	0.57
hfl/chinese-roberta-wwm-ext	0.49	0.76	0.81	0.55

Table 5. Performance of BERT models trained on the original combined dataset (CVAT + CVAS + DSAMST) and evaluated on ROCLING 2025 Test set

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.53	0.7	0.84	0.57
freedomking/mc-bert	0.53	**0.75**	**0.77**	**0.59**
hfl/chinese-bert-wwm	0.54	0.72	0.84	0.57
hfl/chinese-macbert-base	0.55	0.73	0.78	**0.59**
hfl/chinese-roberta-wwm-ext	**0.51**	0.72	0.8	0.56

Table 6. Performance of BERT models trained on the original DSAMST dataset and evaluated on ROCLING 2025 Test set

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.53	0.74	**0.77**	**0.59**
freedomking/mc-bert	0.55	0.74	0.81	0.58
hfl/chinese-bert-wwm	0.53	0.74	**0.77**	**0.59**
hfl/chinese-macbert-base	0.53	0.74	**0.77**	**0.59**
hfl/chinese-roberta-wwm-ext	**0.52**	**0.75**	**0.77**	0.58

Table 7. Performance of BERT models trained on Teacher A augmented dataset and evaluated on ROCLING 2025 Test set

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.51	0.77	0.79	0.57
freedomking/mc-bert	0.51	0.77	**0.78**	0.57
hfl/chinese-bert-wwm	0.5	0.77	0.79	**0.58**
hfl/chinese-macbert-base	**0.48**	0.78	0.79	0.57
hfl/chinese-roberta-wwm-ext	0.51	**0.79**	**0.78**	**0.58**

Table 8. Performance of BERT models trained on Teacher B augmented dataset and evaluated on ROCLING 2025 Test set

Ensemble strategy	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
Chinese Bank Ensemble	**0.46**	0.78	**0.74**	**0.63**
Teacher Ensemble	0.49	0.78	0.76	0.61
Combined Ensemble	0.47	**0.79**	**0.74**	0.62
Top-6 Ensemble	**0.46**	**0.79**	0.75	0.62

Table 9. Performance of various ensemble strategies on ROCLING 2025 test set

Team	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑	Overall Rank
CYUT-NLP	**0.46**	0.78	**0.74**	**0.63**	1
TCU	0.46	**0.81**	0.76	0.61	2
ntulaw_	0.5	0.75	0.79	0.59	3
SCU-NLP	0.51	0.76	0.87	0.59	4
Monokeros	0.53	0.76	0.82	0.58	5
Hey Vergil	0.63	0.62	1.01	0.21	6

Table 10. Comparison of results with other groups

4.4 Performance Comparison Across Training Data on ROCLING 2025 Share Task Test Set

We observe from Table 5 and Table 6 that including general texts (CVAT + CVAS) positively impacts model performance. Models trained on CVAT + CVAS + DSAMST (Table 5) achieved lower Valence MAE and higher PCC, with slightly better Arousal metrics, compared to models trained solely on DSAMST (Table 6). These results indicate that even non-medical general texts provide additional linguistic diversity, which improves model generalization on the ROCLING 2025 Test set (Lee et al., 2025), particularly for Valence prediction.

In contrast, we find that models trained with RAG-augmented data (Teacher A and Teacher B pseudo-labeled datasets, Table 7 and Table 8) show only marginal improvements in Valence and Arousal metrics compared to DSAMST-only training. While the performance gains are limited, we note that our proposed data generation and multi-teacher labeling process maintains model performance when transferring to professional medical texts, preserving the target domain's emotional distribution and language style. These results demonstrate that our augmentation pipeline reliably produces high-quality medical texts, ensuring stable cross-domain adaptation.

Table 9 summarizes the results of four ensemble strategies on the ROCLING 2025 Test set:

- Chinese Bank Ensemble: integrates predictions from all models trained on general texts (Table 5 and Table 6) via weighted averaging to enhance stability.

- Teacher Ensemble: integrates predictions from all models trained on Teacher A/B RAG-augmented datasets (Table 7 and Table 8) via weighted averaging.

- Combined Ensemble: merges predictions from all models in Table 5–8, leveraging both general and RAG-augmented data to improve stability.

- Top-6 Ensemble: selects the six best-performing predictions from Table 5–6 (general text models) and six from Table 7–8 (RAG-augmented models), combining these 12 sets via weighted averaging to maximize complementary information and overall performance.

We analyze the trends and find that the Top-6 Ensemble and Chinese Bank Ensemble achieve the best performance, effectively improving stability and emotion prediction. We observe that the Combined Ensemble performs moderately, slightly affected by weaker predictions, while the Teacher Ensemble shows the lowest performance among the four but still outperforms single models.

Compared with the best-performing baseline (Table 5, CVAT+CVAS+DSAMST), our Top-6 Ensemble (Table 9) achieves an absolute improvement of 0.02 in Valence PCC (0.77 vs. 0.79) and a reduction of 0.03 in Valence MAE (0.49 vs. 0.46). Similarly, for Arousal, PCC improves by 0.03 (0.59 vs. 0.62) while MAE decreases by 0.04 (0.78 vs. 0.74).

Overall, we conclude that the selective ensemble of high-performing predictions is the most effective strategy for enhancing emotion prediction stability and performance.

Moreover, according to the results in Table 10, our system achieved the top-ranked performance in the ROCLING 2025 Shared Task, further validating the effectiveness of the proposed framework.

5 Conclusion

In this study, we propose a three-stage data augmentation framework combining Retrieval-Augmented Generation (RAG) and Multi-Teacher Pseudo-Labeling to enhance Valence–Arousal prediction on physicians' self-reflection texts.

We observe that integrating general texts (CVAT + CVAS) improves model performance, particularly for Valence. Meanwhile, models trained on RAG-augmented datasets maintain stable predictions when transferring to professional medical domains. In our framework, we systematically generate high-quality, domain-consistent synthetic data, leverage multiple teacher models to reduce labeling bias, and filter unreliable samples to ensure dataset quality.

We find that ensemble strategies, especially the Top-6 and Chinese Bank ensembles, further enhance stability and accuracy.

Importantly, this framework is theoretically applicable to other domains, offering a generalizable approach for cross-domain adaptation in low-resource dimensional sentiment analysis.

6 Future Work

In future work, we plan to integrate reinforcement learning (RL) into our framework to optimize the teacher-model architecture and data augmentation pipeline, to guide sample selection, teacher prediction weighting, and to identify reliable augmented data.

We also aim to further enhance our framework through advanced teacher aggregation strategies, improved retrieval methods, and semi-supervised learning, which may improve the quality of augmented datasets and downstream model performance.

Acknowledgments

This study was supported by the National Science and Technology Council under the grant number NSTC 114-2221-E-324-006.

References

Cheng, Yu-Ya, Yan-Ming Chen, Wen-Chao Yeh, and Yung-Chun Chang. 2021. Valence and Arousal-Infused Bi-Directional LSTM for Sentiment Analysis of Government Social Media Management. Applied Sciences, 11(2):880.

Ekman, Paul. 1992. An argument for basic emotions. Cognition & Emotion, 6(3-4):169–200. https://doi.org/10.1080/02699939208411068

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, et al. 2023. Survey of hallucination in natural language generation. ACM Computing Surveys (CSUR), 55(12):1–38. https://doi.org/10.1145/3571730

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33:9459–9474.

Lee, Lung-Hao, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: Building Valence-Arousal Resources for Dimensional Sentiment Analysis. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(4), article 65.

Lung-Hao Lee, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. 2025. ROCLING-2025 Shared Task: Chinese Dimensional Sentiment Analysis for Medical Self-Reflection Texts. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing.

Nguyen, Huy Thong, En-Hung Chu, Lenord Melvix, Jazon Jiao, Chunglin Wen, and Benjamin Louie. 2024. Heuristic-Free Multi-Teacher Learning. arXiv preprint arXiv:2411.12724.

Russell, James A. 1980. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161. https://doi.org/10.1037/h0077714

Schouten, Kim, and Flavius Frasincar. 2015. Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 28(3):813–830. https://doi.org/10.1109/TKDE.2015.2485209

Wang, Jin, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016a. Community-based weighted graph model for valence-arousal prediction of affective words. In Proc. of ROCLING 2021, IEEE/ACM Trans. Audio, Speech and Language Processing, 24(11):1957-1968. https://rocling2021.github.io/

Wang, Jin, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016b. Locally weighted linear regression for cross-lingual valence-arousal prediction of affective words. Neurocomputing, 194:271–278. https://doi.org/10.1016/j.neucom.2016.02.057

Wang, Jin, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2019. Investigating Dynamic Routing in Tree-Structured LSTM for Sentiment Analysis. In Proc. of EMNLP/IJCNLP-19, pages 3423–3428.

Wang, Jin, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2020. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:581–591. https://doi.org/10.1109/TASLP.2019.2959251

Wei, Wen-Li, Chung-Hsien Wu, and Jen-Chun Lin. 2011. A regression approach to affective rating of Chinese words from ANEW. In Proc. of ACII-11, pages 121–131.

Wei, Jason, and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6383–6389. https://doi.org/10.48550/arXiv.1901.11196

Wu, Chuhan, Fangzhao Wu, Yongfeng Huang, Sixing Wu, and Zhigang Yuan. 2017. THU_NGN at IJCNLP-2017 Task 2: Dimensional Sentiment Analysis for Chinese Phrases with Deep LSTM. In Proc. of IJCNLP-17, Shared Tasks, pages 47–52.

Yu, Liang-Chih, Lung-Hao Lee, Jin Wang, and KamFai Wong. 2017. IJCNLP-2017 Task 2: Dimensional Sentiment Analysis for Chinese Phrases. In Proc. of IJCNLP-17, Shared Tasks, pages 9–16.

Yu, Liang-Chih; Wang, Jin; Peng, Bo; Huang, Chu-Ren. 2021. *ROCLING-2021 Shared Task: Dimensional Sentiment Analysis for Educational Texts*. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 385-388. Taoyuan, Taiwan. ACLCLP. URL: https://aclanthology.org/2021.rocling-1.51

NTULAW at ROCLING-2025 Shared Task: Domain-Adaptive Modeling of Implicit Emotions in Medical Reflections

Sieh-chuen Huang
College of Law,
National Taiwan University,
No. 1, Sec. 4, Roosevelt Rd.,
Taipei, Taiwan
schhuang@ntu.edu.tw

Hsuan-Lei Shao*
Graduate Institute of Health
and Biotechnology Law,
Taipei Medical University
301 Yuantong Rd. Zhonghe Dist.
New Taipei City, Taiwan, 235603
hlshao@tmu.edu.tw

Abstract

This paper describes the NTULAW team’s participation in the ROCLING 2025 Dimensional Sentiment Analysis (DSA) shared task, which focuses on predicting valence and arousal ratings for Chinese doctors’ self-reflection texts. Unlike previous editions of the DSA task that targeted words, phrases, or educational comments, this year’s dataset consists of domain-specific multi-sentence medical narratives, posing challenges such as low-arousal writing styles, implicit emotion expressions, and discourse complexity. To address the domain shift between general affective resources (Chinese EmoBank) and medical reflections, we designed a multi-scale BERT-based architecture and explored different data selection strategies. Our final system adopted a hybrid submission: using a model trained solely on doctors’ annotations for arousal prediction, and a combined model with Chinese EmoBank for valence prediction. The system achieved stable performance, ranking third among six participating teams. Error analysis shows systematic overestimation of implicit or negated expressions for valence and regression toward mid-range predictions for arousal. We conclude with limitations of relying only on BERT and outline future work involving domain adaptation, discourse-aware modeling, and large language models (LLMs).

Keywords: Dimensional Sentiment Analysis, Valence—Arousal BERT Modeling, Chinese EmoBank, Medical Self-Reflection Texts, Domain Adaptation

1 Introduction

Sentiment analysis has become one of the most widely studied topics in natural language processing (NLP), with applications ranging from social media mining to healthcare. While categorical sentiment classification maps texts into discrete classes such as positive, negative, or neutral, dimensional sentiment analysis (DSA) offers a more fine-grained representation by positioning affective states in the valence—arousal (VA) space (Russell, 1980). Valence measures the degree of pleasantness, whereas arousal reflects the level of activation from calm to excited. This continuous framework allows researchers to capture subtle affective differences beyond simple polarity.

In Chinese NLP, DSA has been advanced through several shared tasks. The first edition, organized at IALP 2016, focused on word-level prediction (Yu et al., 2016). The newer edition of Chinese EmoBank as a resource for phrase- and sentence-level prediction, educational self-evaluation comments (Lee et al., 2019, 2022). These efforts provided annotated corpora, baselines, and evaluation protocols, which laid the groundwork for subsequent research in this area.

The ROCLING 2025 shared task further extends this line of research into a new and challenging domain: Chinese doctors’ self-reflection writings. Unlike short words or sentences, these multi-sentence texts combine clinical event descriptions with professional reflections(Lee et al., 2025).

In this paper, we present the NTULAW team’s system and findings. We first analyze the differences between Chinese EmoBank and the doctors’ corpus, showing that domain shift is a major factor affecting model performance. We then describe our multi-scale BERT-based architecture and alternative data selection strategies. Our final hybrid submission, which combines domain-specific training for arousal with EmoBank resources for valence, achieved third place among six teams.

*Corresponding author.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 390–398
November 20-22, 2025 ©2025 Association for Computational Linguistics

Finally, we provide quantitative and qualitative error analysis, highlighting how implicit negativity, negation, and mixed polarity remain key challenges for future DSA systems.

2 Related Work

2.1 Dimensional Sentiment Analysis

Dimensional sentiment analysis (DSA) models emotions as continuous values—typically in the valence—arousal (VA) space—providing a finer-grained representation of affective meaning beyond categorical sentiment classification (Russell, 1980). In Chinese NLP, research has explored character-level affective annotations (Peng et al., 2024), hybrid deep learning models such as CNN—BiLSTM for text classification (Liu, 2024), and valence—arousal predictors that combine knowledge-based and embedding-based methods, which ranked top in the IALP 2016 shared task (Wang and Ma, 2016). These efforts demonstrate both the feasibility and effectiveness of applying DSA in the Chinese context (Yu et al.,2016)(Lee et al., 2022).

Recently, DSA research has advanced rapidly with the rise of aspect-based sentiment analysis (ABSA) and transformer-based architectures. The SIGHAN-2024 shared task introduced Chinese Dimensional ABSA, integrating BERT and large language models (LLMs) for entity extraction, relation classification, and intensity prediction (Xu et al., 2024; Lee et al., 2024). A bibliometric review covering 2010—2025 (Gao et al., 2025) revealed a surge in DSA-related studies after 2019, driven by deep learning methods such as BiLSTM, CRF, and attention mechanisms. Further evaluations of transformer variants—BERT, RoBERTa, DistilBERT, and particularly Electra—have demonstrated superior performance in large-scale sentiment classification (Supal et al., 2025). Complementary hybrid approaches combining traditional machine learning and deep neural encoders have improved feature extraction and emotional intensity regression (Singh et al., 2025).

In applied domains, education and public health studies have shown that DSA can capture nuanced affective shifts, such as declining positivity during the post-pandemic transition to in-person learning (Tanquis et al., 2025) and optimistic sentiment toward policy relaxation (Wang and Wang, 2023). However, persistent challenges remain, including data sparsity, class imbalance, and the need for standardized datasets and emotion-expression benchmarks (Yan and Cui, 2025; Kastrati et al., 2021). These recent works collectively underscore the ongoing shift toward transformer-driven, multi-domain approaches to dimensional sentiment modeling—an evolution that also motivates our multi-scale BERT design for medical reflective texts.

2.2 Applications in Chinese NLP

Beyond research on affective resources, sentiment analysis techniques have also been applied across diverse Chinese-language domains, such as e-commerce product reviews (Lee et al., 2019). These studies highlight the practical importance of sentiment analysis in real-world applications while also demonstrating the adaptability of advanced neural architectures.

2.3 Cross-Lingual and Multilingual Perspectives

Sentiment analysis is also an active area in multilingual contexts, where training data in low-resource languages is often generated using machine translation. Comparative experiments have shown that multilingual sentiment analysis with translation-based methods can reach performance comparable to English when combined with supervised learning algorithms (Balahur and Turchi, 2014, 2012a,b). In addition, lexicon construction remains a crucial component, such as the development of sentiment lexicons for Urdu, Roman Urdu, Pashto, and Roman Pashto (Khan et al., 2024). Studies on morphologically rich languages like Arabic have further emphasized the need for comprehensive lexicons and hybrid learning approaches to handle linguistic complexity (Sabih et al., 2018; Obaidat et al., 2015).

2.4 Affective Analysis in the Medical Domain

Despite progress in general and multilingual settings, affective analysis in healthcare remains relatively underexplored. Medical reflective writings contain implicit affective expressions tied to professional experiences and wellbeing. However, few prior studies have addressed dimensional sentiment prediction in this domain, particularly for Chinese. The ROCLING 2025 shared task therefore extends previous work to multi-sentence doctors' self-reflection texts, introducing new challenges in domain-specific language and discourse-level affective modeling.

These gaps in prior work motivate our participation in the ROCLING 2025 shared task, where we specifically address dimensional sentiment prediction in doctors' reflective writings.

3 Research Design

3.1 Task Briefing

The shared task organizers provided two datasets: (1) **Chinese EmoBank**, a general-purpose affective resource for dimensional sentiment analysis, and (2) a domain-specific **validation set** consisting of doctors' self-reflection texts. These two corpora differ significantly in their distributions, sources, and linguistic styles.

Chinese EmoBank. Chinese EmoBank is a large-scale affective resource developed for dimensional sentiment analysis, where each unit (word, phrase, or sentence) is annotated with valence—arousal (VA) scores. It includes several sub-corpora: CVAW (words), CVAP (phrases), CVAT (texts such as reviews), and CVAS (social media posts such as Twitter). The corpus covers multiple genres, ranging from formal written text to colloquial and user-generated content. This diversity results in a broad coverage of affective expressions, including both high-arousal and low-arousal emotions.

Validation Set. The validation set for ROCLING 2025 consists of Chinese doctors' self-reflection writings. These texts are typically multi-sentence narratives describing clinical experiences and professional feelings. Unlike EmoBank, the domain-specific nature of the validation set makes it more homogeneous, focusing on reflective and observational content rather than overtly emotional language. This dataset better represents the task's real-world application in medical contexts.

3.2 Dataset Comparison

We conducted a preliminary comparison between Chinese EmoBank and the doctors' reflection corpus, focusing on their valence—arousal distributions, source characteristics, and stylistic properties.

1. Arousal Distribution. Figure 1–4 illustrates the differences in valence-arousal distribution. Chinese EmoBank (especially CVAW, CVAP, and CVAT) shows a wide spread across the arousal scale, with substantial samples in the mid-to-high range (5—8). In contrast, the doctors' corpus is concentrated in the low-to-mid arousal region (1—7), with very few high-arousal instances. This reflects the writing conventions of clinical texts, where doctors favor neutral and professional language, avoiding overly emotional expressions.

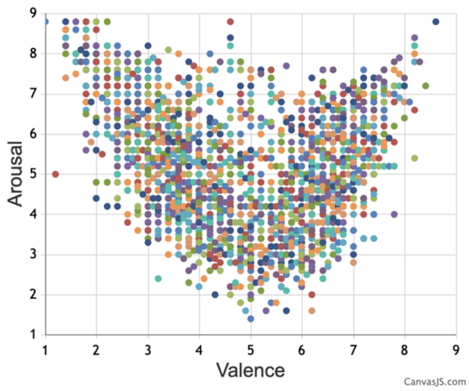


Figure 1: Valence—Arousal distribution of CVAW subset.

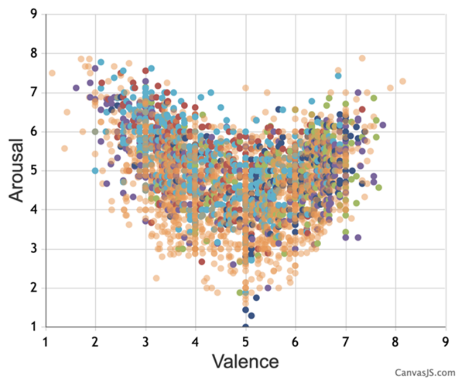


Figure 2: Valence—Arousal distribution of CVAP subset.

2. Source and Diversity. Chinese EmoBank draws from heterogeneous sources,

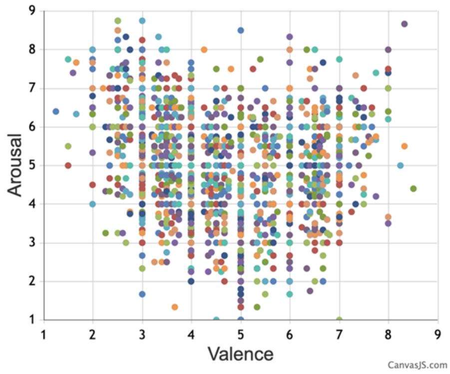


Figure 3: Valence—Arousal distribution of CVAT subset.

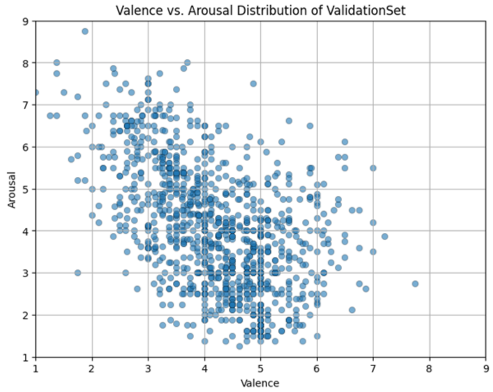


Figure 4: Valence—Arousal distribution of the doctors' validation set.

including news articles, online forums, product reviews (books, hotels, laptops), and Twitter posts. These varied genres contribute to richer emotional content and a wider valence—arousal coverage. In contrast, the doctors' corpus originates from a single domain—medical reflections—focused on clinical records and interactions. This lack of source diversity reduces the presence of explicit affective vocabulary.

3. Style and Emotional Expression. Doctors' reflections typically combine event descriptions with clinical observations, resulting in longer sentences and more complex structures, but with more implicit emotional markers. For example, a sentence like "矛盾心情全透露在他們臉上" (Their conflicting feelings were fully revealed on their faces.) conveys emotion indirectly through observation rather than direct affective terms. In contrast, Chinese EmoBank contains abundant explicit emotion words such as "快樂" (happy), "氣死" (furious), or "害怕" (afraid). Consequently, the doctors' corpus tends to cluster in the low-to-mid valence range with relatively low arousal values.

3.3 Data Selection and Model Design

Our preliminary comparison indicates that a model trained solely on Chinese EmoBank may not accurately capture the characteristics of the doctors' corpus, since medical reflective texts tend to exhibit low-arousal and indirect emotional expressions. To address this issue, we consider **multi-source data integration**, where EmoBank provides general affective knowledge and the doctors' corpus supplies domain-specific adaptation. From a data distribution perspective, domain shift is a critical factor influencing prediction performance, making domain adaptation or multi-task learning necessary for this task.

Training Data and Annotation.

- **Arousal labels**: annotated by medical professionals on the doctors' corpus.
- **Valence labels**: derived from both Chinese EmoBank and doctors' annotations.

Data Selection Strategies.

- **Only-train (Arousal-oriented)**: use only the doctors' annotated data with more reliable arousal labels, ensuring stability and quality for arousal prediction.
- **Train + ChineseEmo (Valence-oriented)**: combine doctors' data with Chinese EmoBank to improve valence prediction, leveraging the richer coverage of valence annotations.

Data Characteristics. Doctors' reflective texts are generally neutral in tone, but contain subtle lexical variations that encode fine-grained emotions. Therefore, models need to be sensitive to weak affective signals while avoiding overfitting to overtly emotional vocabulary found in EmoBank.

3.4 Model Architecture

The proposed system is designed to capture emotional cues in Chinese texts at multiple levels of granularity. Since doctors' reflective writings often encode emotions subtly, we adopt a multi-scale architecture that processes inputs at the sentence, phrase, and word levels. This design allows the model to combine global semantics, local collocations, and character-level nuances, thereby enhancing its sensitivity to weak affective signals.

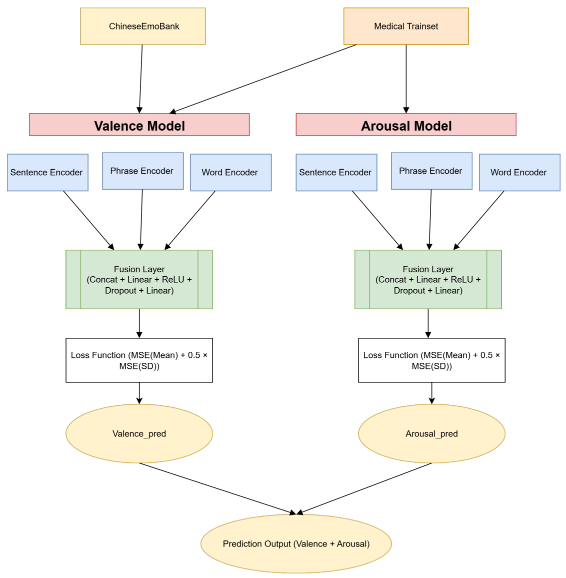


Figure 5: Modal Structure and Processing

Model Input. Three input representations are constructed from each text. At the sentence level, full sentences or paragraphs are used to model discourse and context. At the phrase level, we extract 2-gram segments to highlight collocations and local emotional patterns. Finally, at the word level, character-based sequences are included to capture fine-grained lexical information, which is particularly suitable for Chinese.

Encoder and Fusion Layers. Each input representation is processed by a dedicated BERT encoder. The sentence encoder focuses on capturing global semantics and discourse-level information. The phrase encoder emphasizes phrase-level collocations that often carry affective meaning. The word encoder specializes in character-level features, improving recognition of subtle emotion expressions. Together, the three encoders provide a layered semantic understanding of the input. The output vectors of the three encoders are concatenated and passed through a feed-forward network. This network consists of a linear layer, followed by ReLU activation, dropout for regularization, and a final linear layer that produces two regression outputs: one for arousal and one for valence.

Training Target and Loss. Each text in the dataset is annotated with both mean and standard deviation (SD) values for arousal and valence. The mean ratings ($Arousal_{Mean}$, $Valence_{Mean}$) capture the central tendency of annotators, while the SD values ($Arousal_{SD}$, $Valence_{SD}$) reflect the degree of agreement or variability across annotators. To leverage this richer annotation scheme, we design a composite loss that considers both accuracy and stability. Formally, the loss function is defined as:

$$\begin{aligned}\mathcal{L} = & MSE(\hat{A}, A_{Mean}) + 0.5 \times MSE(\hat{A}, A_{SD}) \\ & + MSE(\hat{V}, V_{Mean}) + 0.5 \times MSE(\hat{V}, V_{SD})\end{aligned} \tag{1}$$

where $\hat{A}$ and $\hat{V}$ denote the predicted arousal and valence scores, respectively. This design encourages the model not only to approximate the average affective ratings but also to account for annotator disagreement, leading to smoother and more robust predictions.

Training Characteristics. We optimize the model using the AdamW optimizer with a linear learning rate scheduler. Early stopping is applied to prevent overfitting and ensure the best performance on the validation set. The architecture also supports cases where phrase-only data are available by inserting dummy vectors for missing sentence or word inputs, ensuring flexibility across different text granularities.

Advantages. The proposed architecture offers three key advantages. First, the multi-scale fusion enables the model to simultaneously capture global context, local collocations, and character-level nuances. Second, the joint prediction of arousal and valence reduces the need for training separate models, making the system more resource-efficient. Finally, by incorporating SD values into the loss, the model becomes more robust to annotator disagreement and produces smoother predictions that are well suited for Chinese reflective texts.

4 Results

4.1 Internal Experiments

We first attempted to build a model trained only on Chinese EmoBank (**M(ChineseE)**). However, the results were unsatisfactory, with high error and weak correlations: MAE of 1.10 and PCC of 0.44 for arousal, and MAE of 0.61 and PCC of 0.65 for valence. This indicates that the model trained purely on general-domain data struggles to adapt to the characteristics of medical reflective texts.

To address this limitation, we explored two alternative data selection strategies. The first, **M(Val)**, used only the valence-annotated subset of the doctors' corpus. This approach achieved the best performance for arousal prediction (MAE = 0.79, PCC = 0.59) and also produced strong results for valence (MAE = 0.50, PCC = 0.72). These findings suggest that restricting training to high-quality annotations enhances prediction stability, particularly for the arousal dimension.

The second approach, **M(Val+ChineseE)**, combined the doctors' valence-annotated data with Chinese EmoBank. This strategy did not improve arousal performance (MAE = 1.10, PCC = 0.44), but slightly enhanced valence prediction (MAE = 0.50, PCC = 0.75) compared to M(Val). This result highlights a trade-off: while external resources enrich valence prediction by providing broader coverage, they may introduce domain shift that harms arousal prediction.

Overall, the experiments demonstrate that domain-specific annotations are crucial for accurate arousal prediction, whereas valence prediction can benefit from multi-source integration with general-domain affective resources. Therefore, in our final submission, we adopted a hybrid approach: using **M(Val)** for arousal prediction and **M(Val+ChineseE)** for valence prediction.

Model	A (MAE,PCC)	V (MAE,PCC)
M(Val)	(0.79, 0.59)	(0.50, 0.72)
M(Val+ChiE)	(1.10, 0.44)	(0.50, 0.75)
M(ChiE)	(1.10, 0.44)	(0.61, 0.65)

Table 1: Internal experiment results with different training data strategies.

4.2 Official Evaluation Results

Table 2 shows the official evaluation results of the shared task. Our system (**ntulaw__**) achieved a balanced performance across all metrics. In particular, the model produced competitive results for both valence and arousal, although slightly behind the top two teams. Overall, our submission ranked **third place** among six participating teams, demonstrating stable and reliable performance.

Team (ID)	V-MAE	V-PCC	A-MAE	A-PCC
CYUT-NLP	0.46	0.78	0.74	0.63
TCU	0.46	0.81	0.76	0.61
ntulaw	0.50	0.75	0.79	0.59
SCU-NLP	0.51	0.76	0.87	0.59
Monokeros	0.53	0.76	0.82	0.58
Hey Vergil	0.63	0.62	1.01	0.21

Table 2: Official evaluation results of the task

5 Discussion

5.1 Error Analysis

To further evaluate the distributional properties of our predictions, we examined Q–Q and P–P plots for valence and arousal.

Valence. As shown in Figure 6, the plots demonstrate a reasonably good alignment along the 45-degree line, though with slight deviations in the mid-to-high quantile range. This suggests that the model captures the central tendency of valence effectively, but tends to underestimate extreme positive values. The P–P plot confirms this observation, showing strong overall agreement between the cumulative distributions of predictions and ground truth.

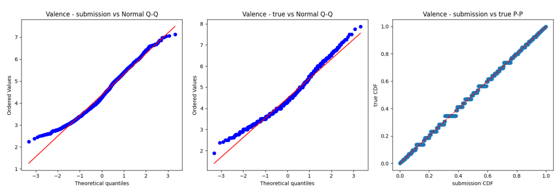


Figure 6: Diagnostic Q–Q and P–P plots for valence predictions.

Arousal. Figure 7 presents similar analyses for arousal. Compared to valence, the predicted arousal values deviate more at the tails, indicating that the model underestimates variance and struggles with extreme arousal levels. Nevertheless, the P–P plot shows that the predicted distribution still closely follows the true cumulative distribution, confirming that the model is reliable in the mid-range but less accurate for highly activated emotional expressions.

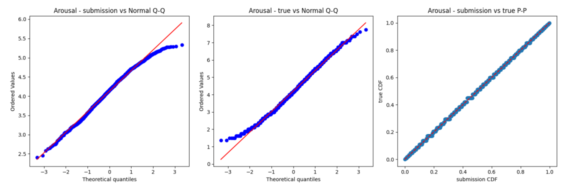


Figure 7: Diagnostic Q–Q and P–P plots for arousal predictions.

5.2 Qualitative Error Analysis (Valence)

We further investigated the sentences with the largest prediction errors in valence. In all cases, the model predicted values much higher than the ground truth, revealing systematic overestimation when emotional expressions are implicit, negated, or embedded in mixed contexts.

For example, in ID V0256 the true valence was 2.63, but the model predicted 5.86. The text contains words like ” 理解”(understanding) and ” 同理”(empathy), which are usually positive in general-domain corpora. Yet here the doctor is empathizing with distressed family members, and the overall emotional tone should remain negative. The lexical cues thus misled the model into an overly positive prediction.

In ID V0013 (true = 1.63, pred = 4.67), the doctor narrates a case in procedural detail, describing responsibility for a critically ill patient. Although emotionally heavy, the text lacks overt negative emotion words, causing the model to assign a mid-range valence.

ID V0185 shows a similar issue: the true rating was 1.25, but the prediction reached 4.18. The sentence reflects on ” 活著”(being alive) while also describing unrelieved pain and swelling. Abstract reflective terms (e.g., 思考, thinking) appear neutral or even positive to the model, diluting the strong negative context of suffering.

Finally, in ID V0064 (true = 2.38, pred = 4.93), the sentence includes the word ” 悲傷”(sadness), but it is negated by ” 不能夠顯露出悲傷的情緒”(Unable to display sorrowful emotions.). The model likely ignored the scope of negation and misinterpreted ” 悲傷”(sadness) as a direct negative marker, again resulting in overestimation.

In summary, these error cases reveal that the model often fails to handle (1) lexical—context mismatches (empathetic words in tragic situations), (2) implicit negativity in reflective writing, (3) mixed polarity within single sentences, and (4) negation scope. Future work should therefore integrate domain-adaptive fine-tuning, negation-aware processing, and discourse-level segmentation to better capture subtle emotional signals in medical reflections.

5.3 Qualitative Error Analysis (Arousal)

We also examined the cases with the largest prediction errors for arousal. In these examples, the model systematically underestimated high-arousal texts and overestimated low-arousal ones, reflecting its tendency to regress towards the mid-range values (around 4—5).

For instance, in ID V0574 the true arousal was 7.88, but the model predicted only 4.45. The sentence describes the final day of ICU training, with urgency and emotional weight. However, the reflective and narrative style downplayed explicit high-arousal cues, leading the model to underestimate the intensity.

A similar pattern appears in ID V0034 (true = 7.50, pred = 4.38). The doctor urgently describes controlling seizures with medications and intubation due to respiratory acidosis. Although the clinical situation is clearly intense, the text contains mostly procedural terms (BZD, Keppra, intubation) that the model may associate with neutral reporting, resulting in lowered arousal prediction.

On the opposite end, ID V0314 (true = 1.38, pred = 4.47) was substantially overestimated. The sentence emphasizes acceptance of illness and appreciation of care, which should indicate calmness. Yet phrases like ” 病人剛好跳短暫 VT”(The patient suddenly went into a short episode of ventricular tachycardia.) introduce suddenness that may have been misinterpreted as high arousal.

Finally, ID V0151 (true = 1.88, pred = 4.93) illustrates reflective calmness: "我們是不是可以更加專注做自己想做的事情了.(Can we now focus more on doing what we truly want to do?)" This expresses philosophical contemplation rather than excitement. Nevertheless, the rhetorical framing and modal verb usage might have been interpreted as emotionally charged, causing overestimation.

In summary, the arousal errors reveal two major tendencies: (1) underestimation of truly high-arousal emergency contexts, when described with technical or reflective wording; and (2) overestimation of calm or philosophical passages that include interrogatives, sudden events, or modal expressions. Future work should incorporate domain-adaptive embeddings that better distinguish between clinical urgency and rhetorical style, as well as discourse-level modeling to capture shifts between calm reflection and acute events.

6 Conclusion

6.1 Overall Task Review

In this paper, we presented our system for the ROCLING 2025 Dimensional Sentiment Analysis (DSA) shared task, focusing on doctors' reflective texts. Our study highlighted the challenges of applying general-domain affective resources, such as Chinese EmoBank, to the medical domain. Through systematic experiments, we found that domain-specific annotations are crucial for arousal prediction, whereas valence prediction benefits from multi-source integration. Based on these findings, we adopted a hybrid submission strategy: using the M(Val) model for arousal and the M(Val+ChineseE) model for valence. This approach achieved stable performance, ranking third among six participating teams.

Beyond quantitative results, our qualitative error analysis revealed important insights into model limitations. For valence, errors often stemmed from empathetic words used in tragic contexts, implicit negativity, mixed polarity, and negation scope. For arousal, the model underestimated high-arousal emergency descriptions that were written in technical terms, and overestimated calm, reflective passages that contained interrogatives or rhetorical devices.

6.2 Limitations

We outline the main limitations of our work and discuss directions for future improvements.

Looking forward, we plan to incorporate domain-adaptive fine-tuning, negation- and discourse-aware modeling, and clause-level segmentation to better capture subtle emotional signals. In addition, variance-aware training objectives may help the system better model extreme values on both valence and arousal scales.

Another limitation of this study is that our experiments relied mainly on BERT-based encoders. While effective, such models may lack the capacity to fully capture nuanced discourse and implicit affective cues. Future work should therefore explore larger pre-trained language models (LLMs) and hybrid architectures, combining domain adaptation, variance-aware objectives, and sentence-level reasoning to better capture subtle emotional signals.

By participating in ROCLING 2025, we aimed to bridge computational linguistics, healthcare, and law—demonstrating how interdisciplinary collaboration can contribute to affective computing research in Chinese NLP.

Ethical Considerations

The dataset used in this study was provided by the ROCLING 2025 shared task organizers and consists of anonymized Chinese doctors' self-reflection texts. No personally identifiable information (PII) was included, and we did not conduct any additional data collection. Our models are intended solely for research purposes. They should not be applied directly in clinical decision-making, as misinterpretation of affective predictions in sensitive medical contexts may pose ethical risks.

References

A. Balahur and M. Turchi. 2012a. Comparative experiments for multilingual sentiment analysis using machine translation. In *CEUR Workshop Proceedings*.

A. Balahur and M. Turchi. 2012b. Multilingual sentiment analysis using machine translation? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

A. Balahur and M. Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech and Language*, 28(1):56–75.

R. Gao, Y. Liu, and Y. Qiao. 2025. Trends and developments in aspect-based sentiment analysis: A bibliometric study using citespace and the web of science database (2010–2025). In *Proceedings of the 4th International Symposium on Computer Applications and Information Technology (ISCAIT 2025)*. IEEE.

Z. Kastrati, F. Dalipi, A. S. Imran, and M. A. Wani. 2021. Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. *Applied Sciences (Switzerland)*, 11(3):987–1002.

Z. A. Khan, Y. Xia, A. Khan, and E. A. A. Ismail. 2024. Developing lexicons for enhanced sentiment analysis in software engineering: An innovative multilingual approach for social media reviews. *Computers, Materials and Continua*.

J. S. Lee, D. Zuba, and Y. Pang. 2019. Sentiment analysis of chinese product reviews using gated recurrent unit. In *Proceedings of the 5th IEEE International Conference on Big Data Service and Applications (BigDataService)*.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(4):Article 65.

Lung-Hao Lee, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. 2025. Rocling-2025 shared task: Chinese dimensional sentiment analysis for medical self-reflection texts. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*.

Lung-Hao Lee, Liang-Chih Yu, S. Wang, and J. Liao. 2024. Overview of the sighan-2024 shared task for chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics.

Y. Liu. 2024. Role of natural language processing in document understanding and semantic analysis: A chinese perspective. *Profesional de la Informacion*.

I. Obaidat, R. Mohawesh, M. Al-Ayyoub, and Y. Jararweh. 2015. Enhancing the determination of aspect categories and their polarities in arabic reviews using lexicon-based approaches. In *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*.

C. Peng, X. Xu, and Z. Bao. 2024. Sentiment annotations for 3827 simplified chinese characters. *Behavior Research Methods*.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

S. Sabih, A. Sallam, and G. S. El-Taweel. 2018. Manipulating sentiment analysis challenges in morphological rich languages. In *Advances in Intelligent Systems and Computing*.

B. Singh, K. Kaur, and G. Kaur. 2025. Optimizing emotion detection: An nlp-driven deep learning approach to sentiment encoding. In *Proceedings of the International Conference on Data Science and Business Systems (ICDSBS 2025)*. IEEE.

S. Supal, S. M. Anzar, C. Jacob, and D. Aji. 2025. Deep learning transformers for sentiment classification: A performance evaluation. In *Proceedings of the 6th International Conference on Control Communication and Computing (ICCC 2025)*. IEEE.

J. Q. Tanquis, L. Feliscuzo, and C. L. S. Romana. 2025. Data collection tools in faculty evaluation sentiment analysis. In *Proceedings of the 9th International Symposium on Innovative Approaches in Smart Technologies (ISAS 2025)*. IEEE.

H. Wang and X. Wang. 2023. Sentiment analysis of tweets and government translations: Assessing china's post-covid-19 landscape for signs of withering or booming. *Global Media and China*.

H.-Y. Wang and W.-Y. Ma. 2016. Ckip valence-arousal predictor for ialp 2016 shared task. In *Proceedings of the 2016 International Conference on Asian Language Processing (IALP)*.

H. Xu, D. Zhang, Y. Zhang, and R. Xu. 2024. Hitsz-hlt at sighan-2024 dimabsa task: Integrating bert and llm for chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics.

S. Yan and S. Cui. 2025. Fine-grained sentiment analysis of movie reviews based on machine learning and deep learning models. In *Proceedings of the 5th Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS 2025)*. IEEE.

Liang-Chih Yu, Lung-Hao Lee, and Kam-Fai Wong. 2016. Overview of the ialp 2016 shared task on dimensional sentiment analysis for chinese words. In *Proceedings of the 2016 International Conference on Asian Language Processing (IALP)*, pages 156–160. IEEE.

TCU at ROCLING-2025 Shared Task: Leveraging LLM Embeddings and Ensemble Regression for Chinese Dimensional Sentiment Analysis
運用大型語言模型嵌入與集成迴歸技術進行中文維度情感分析

李信頡 **(Hsin-Chieh Li)**
Department of Medical Informatics
Tzu Chi University
113206105@gms.tcu.edu.tw

林紋正 **(Wen-Cheng Lin)***
Department of Medical Informatics
Tzu Chi University
denislin@gms.tcu.edu.tw

摘要

本文參加 ROCLING-2025 共享任務：針對醫學自我反思文本的中文維度情感分析。維度式情感分析（Dimensional Sentiment Analysis, DSA）將情緒視爲連續維度，如效價（valence，正向至負向）與喚醒度（arousal，平靜至激動），相較傳統分類式方法提供更細膩的表徵，適用於心理健康監測與風險預測等應用。利用大型語言模型（Large Language Models, LLMs）作爲特徵提取器生成上下文嵌入向量，並應用於支持向量迴歸（SVR）等迴歸模型進行效價-喚醒度預測。訓練資料採用 Chinese EmoBank 資料集（2,954 筆通用領域樣本），驗證資料爲醫療自省文件語料資料集（994 筆），測試資料爲醫療自省文件語料資料集（1,541 筆）。實驗結果顯示，使用 DeepSeek 嵌入的 SVR 模型表現最佳，在此基礎上透過多模型集成學習，效能提升至 valence MAE：0.463、arousal MAE：0.759、valence PCC：0.805、arousal PCC：0.608。此方法突顯多模型融合在 DSA 於生醫情境的潛力，促進非侵入式心理健康評估工具的發展。

Abstract

This study participates in the ROCLING-2025 shared task on Chinese dimensional sentiment analysis for medical self-reflection texts. Dimensional Sentiment Analysis (DSA) represents emotions as continuous dimensions, such as valence (positive to negative) and arousal (calm to excited), providing finer-grained representations compared to traditional categorical approaches, which are suitable for applications in mental health monitoring and risk assessment. We use large language models (LLMs) to extract contextual embedding vectors, which are then fed into regression models, such as Support Vector Regression (SVR), for valence-arousal prediction. The training data consists of the Chinese EmoBank dataset (2,954 general-domain samples), the validation data is a Medical Self-Reflection Corpus Dataset (994 samples), and the test data is another Medical Self-Reflection Corpus Dataset (1,541 samples). Experimental results show that the SVR model with DeepSeek embeddings performs best. Multi-model ensemble learning further improves performance to 0.463 valence MAE, 0.759 arousal MAE, 0.805 valence PCC, and 0.608 arousal PCC. This approach shows the potential of multi-model fusion in DSA for biomedical applications, facilitating the development of non-intrusive mental health assessment tools.

關鍵字：維度式情感分析、大型語言模型、集成學習

Keywords: Dimensional Sentiment Analysis, Large Language Models, Ensemble Learning

1 引言

情感分析（Sentiment Analysis），又稱爲意見挖掘（Opinion Mining），是自然語言處理（Natural Language Processing, NLP）領域中的核心任務之一，旨在自動辨識、提取與量化文本資料中所表達的主觀情感與態度。該技術廣泛應用於社群媒體評論分析、產品意見理解、政治輿論監測等領域，並逐漸拓展至醫療與心理健康等應用場景，用以協助病患情緒追蹤與心理風險預測 (Liu, 2012)。

情緒在心理健康評估中扮演關鍵角色，隨著數位語言資料與人工智慧技術的發展，學界與醫療實務者逐漸從日常語言中探索情緒狀態與心理健康的關聯。

研究指出，情緒解析能力（辨識與表達細緻情緒的能力）與心理健康密切相關 (Vishnubhotla et al., 2024)。在社群媒體上，情緒表達模糊或情緒用語貧乏的使用者，更易呈現憂

*通訊作者 (Corresponding Author)

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 399–406
November 20-22, 2025 ©2025 Association for Computational Linguistics

鬱、焦慮等風險，反映情緒語言的豐富程度可作爲非侵入式心理健康指標。

另一方面，線上心理健康社群的語言分析發現，不同精神疾病群體展現出獨特的情緒語言特徵 (Yan et al., 2021)。例如，焦慮群體常帶有未來導向與警戒感，憂鬱群體則偏向無望與自我否定。此差異有助於辨識心理健康問題，且社群互動中的回應亦具情緒調節作用。

綜合而言，情感分析在心理健康預測、疾病風險篩檢與數位健康監控方面展現潛力，透過日常語言的情緒線索，有望發展低成本、即時性的心理健康監測工具，推動預防醫學與個人化照護。

傳統的情感分析多採分類式方法（categorical approach），將情緒分爲「正面」、「負面」或「中性」等類別。然而，這類方法在面對語言中細膩或混合的情緒表達時，往往難以準確反映眞實的心理狀態 (Mohammad, 2016)。爲此，學界提出維度式情感分析（Dimensional Sentiment Analysis, DSA）作爲替代方案，依據心理學理論將情緒視爲連續空間中的向量位置，常見的表示方式包括「效價（Valence）」、「喚醒度（Arousal）」二個維度。(Calvo and D'Mello, 2010; Jonathan Posner and Peterson., 2005)。

ROCLING-2025 舉辦了共享任務：針對醫學自我反思文本的中文維度情感分析 (Lee et al., 2025)，在本任務中參與者需要爲每位醫生的自我反思文本，在情緒效價和喚醒度維度上給出 1 到 9 的實際值評分。本文將說明我們在此任務中使用的方法與結果。

2 相關工作

2.1 早期方法：基於詞典與傳統機器學習

早期研究多依賴於情感詞典與傳統機器學習模型。情感詞典法透過預先標註好 VA 值的情感詞典（如 ANEW, CVAW）(Warriner et al., 2013; Yang et al., 2016)，計算文本中詞語的平均 VA 分數作爲整句的情感預測。傳統機器學習方法，如支持向量迴歸（Support Vector Regression, SVR）(Drucker et al., 1997)，則將文本轉換爲詞袋（Bag-of-Words）或 TF-IDF 等特徵向量，再進行迴歸預測 (Malandrakis et al., 2013)。

2.2 深度學習

以卷積神經網路（CNN）和循環神經網路（RNN）爲代表的模型被廣泛應用於 DSA。CNN 擅長擷取局部特徵 (Kim, 2014)，而 LSTM 等 RNN 變體則能處理序列資訊 (Tai et al., 2015)。有研究將 CNN 與 LSTM 結合，以同時捕捉局部與全局語意 (Hasib et al., 2023)。此外，注意力機制（Attention Mechanism）的引入亦顯著提升了預測效能 (Yang et al., 2016)，(林巍, 2022) 則指出多任務學習有助於捕捉維度間的關聯性。

爲克服傳統「詞袋模型」（Bag-of-Words）無法保留語序與語意的限制，詞嵌入（Word Embedding）技術被提出，用於將文字轉換爲實數向量，以利於語義分析。早期方法如 word2vec (Mikolov et al., 2013) 與 GloVe (Pennington et al., 2014) 爲每個詞生成固定向量；隨著 Transformer 架構的興起，BERT (Devlin et al., 2019) 等預訓練語言模型能提供上下文相關的「語境嵌入」（Contextual Embedding），使同一詞在不同語境中具有不同表示，大幅提升了語意建模能力。

2.3 大型語言模型

大型語言模型 Large Language Models, LLMs）是由具有大量參數的類神經網路組成的一類語言模型，如 GPT (OpenAI, 2023)、Llama (Touvron et al., 2023)、DeepSeek (Guo et al., 2025) 等，展現出驚人的語意理解與生成能力。研究者常以其作爲特徵提取器，將輸出的語意向量輸入傳統機器學習模型（如 SVR、隨機森林）進行 VA 值預測。

2.4 近期方法

首先，在情感維度方面，研究逐漸引入第三維（Dominance / 支配度）或更多維度，以捕捉情緒控制力與社會互動等特徵。Yang 等人提出的 SCCL 模型（Semantic Cluster-level Contrastive Learning）(Yang et al., 2023)，在對話式情緒識別中利用 VAD 三維空間結合對比學習，有效提升模型可解釋性與穩定性。

其次，Transformer 架構與參數高效微調（Parameter-Efficient Fine-Tuning, PEFT）(Vaswani et al., 2023) 技術被廣泛應用於情緒迴歸任務。部分研究透過在預訓練模型上附加迴歸層進行微調，或採用提示式（prompt-based）策略，於少樣本與跨領域任務中取得良好成效 (Wawer, 2024)。

此外，對比學習（contrastive learning）成爲強化情緒嵌入的重要手段。Hu 等人的 LaSCL（Label Semantic-Driven Contrastive Learning）(Hu et al., 2025) 利用情緒標籤語義嵌入作爲語義錨點以強化情緒區分，Xie 等人的 DCLF（Dual Contrastive Learning Framework）(Xie et al., 2025) 則結合多模態與上下文對比學習，以提升特徵融合的整體效果。

在多模態融合方面，近期研究整合文本、語音、影像及生理訊號等多來源特徵，並透過跨模態注意力（cross-attention）與維度專屬融合

機制（dimension-wise fusion）改善模型一致性與穩健性。例如 DCLF 模型與 PCMDA（Parallel Contrastive Multimodal Domain Adaptation）(Li et al., 2025) 均在多模態情緒識別中展現良好成效。

最後，領域適配（domain adaptation）與多指標評估逐漸受到重視。Wawer (Wawer, 2024) 探討少樣本學習於情感分析中的跨域適應，顯示原型網路與相似度式學習對資料稀缺問題具優勢。同時，評估方式亦從單一誤差指標擴展至 MAE、PCC、CCC 等多重指標，以全面反映模型在連續值預測的表現。

3 方法

本研究以大型語言模型所提取之文本特徵作爲輸入，並將其應用於多種機器學習方法進行情緒迴歸任務的建模。所選用之模型涵蓋傳統機器學習方法與深度學習方法兩大類，具體如下：

- 傳統機器學習方法：
  SVR、
  LightGBM (Ke et al., 2017)、
  XGBoost (Chen and Guestrin, 2016)、
  CatBoost (Prokhorenkova et al., 2017)。
- 深度學習方法：
  以多層神經網路爲基礎之迴歸模型。

3.1 資料集

訓練資料集使用 Chinese EmoBank (Lee et al., 2022; Yu et al., 2016) 內容包含 2,954 筆通用領域中文文本，涵蓋新聞、政治、飯店評論、書籍、汽車及筆記型電腦等。文本長度：平均 57.6 字元，標準差 32.6，多數集中於 51–100 字元。情感標註：Valence 平均 4.8，Arousal 平均 4.8，近似常態分佈，集中於中間值。驗證資料集：內容包含 994 筆醫療自省文本，高頻詞爲醫學術語（如「病人」、「治療」）。文本長度：平均 76.5 字元，標準差 53.1，長於訓練集，呈右偏分佈。情感標註：Valence 平均 4.1，Arousal 平均 4.0，情感偏負面且喚醒度較低。測試資料集（DSAMST）：內容包含 1541 筆醫療自省文本。所有資料皆依照維度式情感模型（Dimensional Sentiment Model）進行標註，標籤包含兩個連續變數：

- 情感效價 (Valence)：表示情感的正向或負向程度，評分範圍爲 1 至 9 分，數值越高代表越正面（如愉悅、滿意），數值越低則偏向負面（如悲傷、憤怒）。
- 喚醒度 (Arousal)：表示情感的強度或激動程度，評分範圍同爲 1 至 9 分，數值越高代表越激動或精力充沛（如亢奮、緊張），數值越低則表示較爲平靜或低落（如冷靜、無力）。

3.2 評估指標

模型的性能將根據兩個指標進行評估，分別針對 Valence 和 Arousal 獨立計算:

平均絕對誤差 (Mean Absolute Error, MAE) 評估模型預測值與人工標註眞實值之間的平均絕對差距。其計算公式如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i| \qquad (1)$$

其中，n 爲樣本數，a_i 爲眞實值，p_i 爲預測值。MAE 越小，表示模型誤差越低。

皮爾森相關係數 (Pearson Correlation Coefficient, PCC) 評估模型預測值與眞實值之間的線性相關程度，範圍在-1 到 1 之間。其計算公式如下：

$$PCC = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{a_i - \mu_A}{\sigma_A}\right)\left(\frac{p_i - \mu_P}{\sigma_P}\right) \qquad (2)$$

其中，μ_A 與 μ_P 分別爲眞實值與預測值的平均數，σ_A 與 σ_P 爲其標準差。PCC 越接近 1，表示相關性越強。

3.3 情感分析模型

本研究的核心流程爲：首先使用大型語言模型作爲編碼器（Encoder），將醫學文本轉換爲上下文語意嵌入向量；接著將嵌入向量輸入多種

Combination	MAE_V	MAE_A	PCC_V	PCC_A
Training	0.528	0.914	0.711	0.483
Validation	0.553	0.801	0.706	0.537
Training + half validation	0.524	0.809	0.754	0.538
Training + Validation	0.524	0.809	0.754	0.538

Table 1: 不同訓練資料組合實驗結果

Model Name	MAE_V	MAE_A	PCC_V	PCC_A	Overall Rank
XGBoost	0.526 (3)	0.857 (5)	0.724 (5)	0.457 (5)	4.50 (5)
LGBM	0.519 (1)	0.852 (4)	0.740 (3)	0.472 (4)	3.00 (3)
CatBoost	0.538 (4)	0.834 (3)	0.732 (4)	0.495 (3)	3.50 (4)
SVR	0.524 (2)	0.809 (1)	0.754 (2)	0.538 (2)	1.75 (1)
CustomResNet	0.550 (5)	0.830 (2)	0.761 (1)	0.542 (1)	2.25 (2)

Table 2: 不同機器學習模型之實驗結果比較

Encoder Name	MAE_V	MAE_A	PCC_V	PCC_A
DeepSeek-R1-0528-Qwen3-8B	0.524	0.809	0.754	0.538
DeepSeek-Prover-V1.5-RL	0.527	0.839	0.749	0.504
Llama3-TAIDE-LX-8B-Chat-Alpha1	0.529	0.827	0.743	0.518
multilingual-e5-large	0.554	0.810	0.726	0.531
multilingual-e5-large-instruct	0.523	0.807	0.742	0.539

Table 3: 不同編碼器之實驗結果比較

迴歸模型（Regressors），以預測效價（Valence）與喚醒度（Arousal）數值。每一次實驗（Run）即對應一組「Encoder + Regressor」的組合，並比較其效能表現。

LLM Encoders 本研究使用五種大型語言模型作爲編碼器，以產生輸入文本的上下文嵌入向量。DeepSeek-R1-0528-Qwen3-8B 與 DeepSeek-Prover-V1.5-RL (Xin et al., 2024) 均屬於近年提出的 DeepSeek 系列模型，前者強調通用語意理解，後者則在推理與邏輯任務上進行優化。Llama3-TAIDE-LX-8B-Chat-Alpha1 (TAIDE, 2025) 爲針對中文及多語言對話任務調校的 Llama3 模型版本，具備良好的語境捕捉能力。另一方面，multilingual-e5-large 與 multilingual-e5-large-instruct (Wang et al., 2024) 則爲針對檢索與語意相似度任務設計的嵌入模型，能在多語言環境下提供高效且一致的語意表示。這些編碼器能將輸入文本轉換爲高維度向量表示，進一步保留與情緒相關的語意特徵，並支援後續的迴歸任務。

Regression Models 我們在實驗中比較了多種迴歸模型，涵蓋傳統機器學習方法與深度學習方法。支持向量迴歸（SVR）採用 RBF kernel，主要參數設爲 $C = 10$ 與 $\epsilon = 0.2$，能有效處理非線性分佈，其核心思想是透過高維特徵映射尋找最佳迴歸超平面。XGBoost 模型則使用 $n\_estimators = 1000$、$max\_depth = 6$ 與 $learning\_rate = 0.05$，以 RMSE 作爲評估指標，並透過梯度提升（Gradient Boosting）迭代訓練多棵弱分類樹來提升預測效果。LightGBM 採用 $n\_estimators = 500$、$num\_leaves = 31$ 與 $learning\_rate = 0.05$，其特點是基於葉節點生長策略（Leaf-wise Growth）以提高效率並降低記憶體消耗。CatBoost 模型則設定 $iterations = 1000$、$depth = 6$ 與 $learning\_rate = 0.05$，同樣以 RMSE 作爲損失函數，並利用有序提升（Ordered Boosting）技術減輕資料偏差。

在深度學習方法部分，本研究設計了一個基於殘差結構的多層迴歸模型，底層採用自定義的 Custom ResNet。其架構包含六個全連接殘差模塊，每一模塊由全連接層、均方根正規化（RMSNorm）、ReLU 激活函數與 Dropout（0.5）組成，並透過殘差連接強化深層結構的穩定性與可訓練性。輸入首先經 RMSNorm 正規化，隨後依序通過六個殘差模塊，並由兩層全連接層完成降維與映射，最終輸出長度爲二的向量，分別對應效價（Valence）與喚醒度（Arousal）。在輸出階段，模型引入正弦函數作爲平滑激活，並透過縮放和平移操作將預測值限制於 1 至 9 的範圍，其轉換公式如下：

$$\hat{y} = \sin(x) \times 4 + 5 \qquad (3)$$

其中，x 爲模型線性層的輸出，$\hat{y}$ 爲最終預測分數，確保結果符合任務的標註需求。

3.4 實驗結果

首先，表 1 顯示了不同訓練資料組合的效能表現。本研究的嵌入模型採用 DeepSeek-R1-0528-Qwen3-8B，負責將輸入文本轉換爲高維語意向量；迴歸模型則選用支持向量迴歸（SVR），用以進行情緒維度的連續數值預測。在僅使用訓練集的情況下，Valence 與 Arousal 的 MAE 分別爲 0.528 與 0.914，而 PCC 分別爲 0.711 與 0.483。當額外引入部分驗證集樣本時，模型效能些微改善，最終在

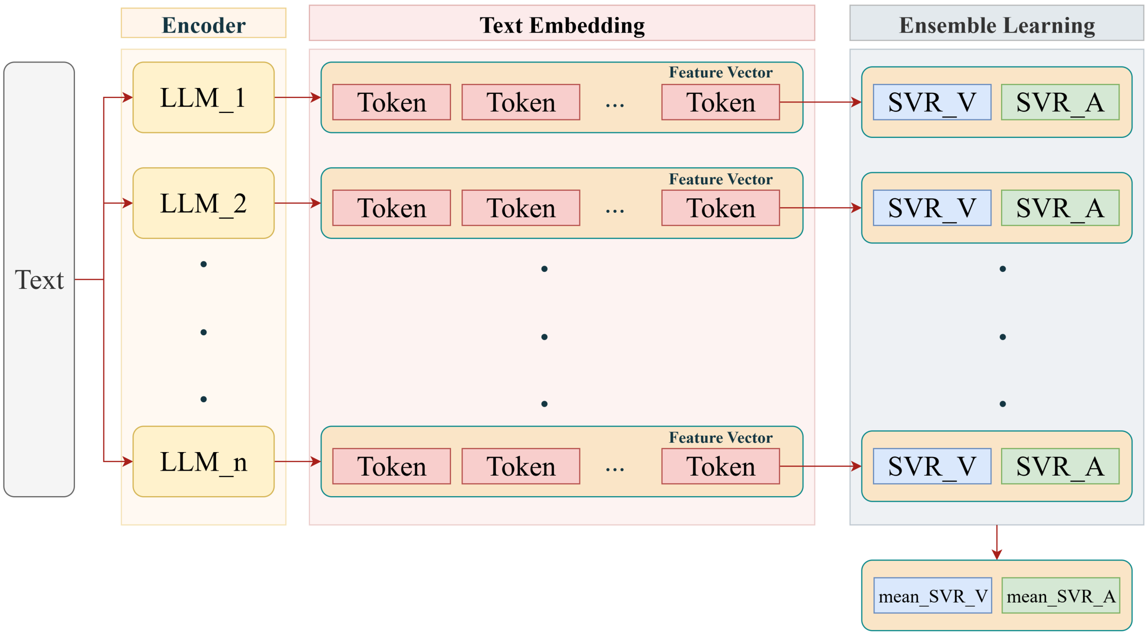


Figure 1: 多編碼器集成學習架構

	MAE_V	MAE_A	PCC_V	PCC_A
Models	0.495	0.802	0.772	0.544
Encoders	0.463	0.759	0.805	0.608

Table 4: 集成式方法

「訓練集 + 驗證集一半」的配置下達到最佳表現，Valence MAE 爲 0.524、Arousal MAE 爲 0.809，對應的 PCC 分別爲 0.754 與 0.538。

進一步比較不同機器學習模型（表 2），本研究在「Training + half validation」訓練策略下，並以 DeepSeek-R1-0528-Qwen3-8B 所生成的嵌入特徵進行實驗。結果顯示，SVR 與 Custom ResNet 的整體表現優於其他梯度提升類模型。SVR 在四項指標上均保持穩定，其中 Valence PCC 爲 0.754，Arousal PCC 爲 0.538，整體排名第一。Custom ResNet 雖在 Valence PCC（0.761）與 Arousal PCC（0.542）上略優，但其 MAE 表現稍遜於 SVR，整體排名第二。這表明 SVR 在小樣本且具跨領域差異的情境中展現出更佳的泛化能力。

在編碼器比較方面（表 3），本研究採用支持向量迴歸（SVR）作爲迴歸方法，訓練資料使用 Training + validation 進行實驗。結果顯示，DeepSeek-R1-0528-Qwen3-8B 與 multilingual-e5-large-instruct 的整體效能相近，其中前者在 Valence PCC（0.754）上取得最佳表現，而後者則在 Arousal PCC（0.539）略優。其他編碼器如 DeepSeek-Prover-V1.5-RL 與 Llama3-TAIDE-LX-8B-Chat-Alpha1 亦展現出一定的穩定性，但整體效能稍低。相較之下，multilingual-e5-large 在四項指標中皆爲最弱，顯示專爲多語言檢索設計的模型在醫療情境下的適配性有限。整體而言，針對中文醫療文本，DeepSeek 系列編碼器在 Valence 與 Arousal 的預測上展現出更佳的穩定性與適配性。

此外，我們使用集成學習方法進一步提升模型效能，並針對不同層級的集成方式進行比較，其結果如表 4 所示。具體而言，Models 爲將表 2（不同迴歸模型）之結果進行集成，即結合 SVR、Custom ResNet 與梯度提升類模型的預測輸出將其取平均值；而 Encoders 則爲將表 3（不同編碼器）之結果進行集成，即融合多個 LLM 編碼器所生成的語意嵌入。比較結果顯示，編碼器層級的集成效果顯著優於模型層級集成，特別是在相關性指標上，Valence PCC 達到 0.805，Arousal PCC 爲 0.608，均高於模型層級集成的表現（Valence PCC = 0.772、Arousal PCC = 0.544）。此結果說明，透過融合不同語意表示能更有效捕捉醫療文本中的情緒訊號，對於 DSA 任務

的效能提升尤爲關鍵。

最後，圖 1 展示了所提出的多編碼器集成架構。該方法使用不同 LLM 編碼器生成的語意嵌入，在迴歸模型預測 V、A 值後取平均值進行融合，以充分利用多視角的語意特徵。實驗結果證實此架構能進一步提升預測準確性，展現了多編碼器集成學習在醫療語境下進行中文維度情感分析的潛力。

3.5 結論

本研究針對 ROCLING-2025 共享任務的中文醫療自我反思文本，系統性地探討了大型語言模型嵌入特徵與多種迴歸方法在維度情感分析任務中的表現，並進行多層次比較與驗證。綜合實驗結果與分析，主要結論如下：

1. 領域適應的重要性：在訓練過程中適度引入目標領域資料（例如採用「Training + half validation」策略），能有效緩解通用語料與醫療文本之間的領域落差，些微提升模型在效價（Valence）與喚醒度（Arousal）預測上的準確度。

2. 迴歸模型比較：在多種迴歸方法中，支持向量迴歸（SVR）與 Custom ResNet 表現最佳。SVR 在少量資料下訓練快速，且超參數需求較低，四項指標整體均衡，展現出穩健的泛化能力；Custom ResNet 則在部分 PCC 指標上略優，但在 MAE 上不及 SVR，整體表現次之。

3. 編碼器效能分析：不同編碼器對情感特徵擷取的能力差異明顯。其中，DeepSeek-R1-0528-Qwen3-8B 在 Valence 預測上達到最佳表現，而 multilingual-e5-large-instruct 則在 Arousal 預測上略佔優勢。整體而言，DeepSeek 系列編碼器在中文醫療文本的情感特徵建模上展現出更高的穩定性與適配性。

4. 集成學習的優勢：相較於單一編碼器或僅在模型層級進行集成，多編碼器的特徵層級集成策略能更有效整合不同模型的優勢，進一步提升模型的泛化能力與穩健性。實驗結果證實，集成架構在 Valence 與 Arousal 的多項指標上均達到最優表現。

綜合而言，本研究證明了將大型語言模型作爲特徵提取器，並結合適當的迴歸方法與集成策略，能有效提升中文維度情感分析在醫療場景中的應用價值。此方法不僅在醫療自我反思文本上展現了優異效能，也爲發展非侵入式、基於自然語言的心理健康評估工具提供了技術支撐與實證基礎。

References

Rafael A. Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37.

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*, pages 785–794.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.

Daya Guo et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Khan Md Hasib, Sami Azam, Asif Karim, Ahmed Al Marouf, F M Javed Mehedi Shamrat, Sidratul Montaha, Kheng Cher Yeo, Mirjam Jonkman, Reda Alhajj, and Jon G. Rokne. 2023. Mcnn-lstm: Combining cnn and lstm to classify multi-class text in imbalanced news data. *IEEE Access*, 11:93048–93063.

Jiaxi Hu, Leyuan Qu, Haoxun Li, and Taihao Li. 2025. Label Semantic-Driven Contrastive Learning for Speech Emotion Recognition. In *Interspeech 2025*, pages 4348–4352.

James A. Russell Jonathan Posner and Bradley S. Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715—734.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(4).

Lung-Hao Lee, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. 2025. Rocling-2025 shared task: Chinese dimensional sentiment analysis for medical self-reflection texts. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Dongdong Li, Shengyao Huang, Li Xie, Zhe Wang, and Jiazhen Xu. 2025. Neuron perception inspired eeg emotion recognition with parallel contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):14049–14062.

Bing Liu. 2012. *Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies*, volume 5. Morgan & Claypool Publishers.

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Saif M. Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Herbert L. Meiselman, editor, *Emotion Measurement*, pages 201–237. Woodhead Publishing.

OpenAI. 2023. Gpt-4 technical report. `https://cdn.openai.com/papers/gpt-4.pdf`.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2017. Catboost: unbiased boosting with categorical features. *arXiv preprint arXiv:1706.09516*.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *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)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

TAIDE. 2025. Llama3-TAIDE-LX-8B-Chat-Alpha1. `https://huggingface.co/taide/Llama3-TAIDE-LX-8B-Chat-Alpha1`. Accessed: 2025-09-01.

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. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Krishnapriya Vishnubhotla, Daniela Teodorescu, Mallory J. Feldman, Kristen A. Lindquist, and Saif M. Mohammad. 2024. Emotion granularity from text: An aggregate-level indicator of mental health.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Amy B. Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Aleksandra Wawer. 2024. Few-shot methods for aspect-level sentiment analysis. *Information*, 15(11):664.

Yunhe Xie, Chengjie Sun, Ziyi Cao, Bingquan Liu, Zhenzhou Ji, Yuanchao Liu, and Lili Shan. 2025. A dual contrastive learning framework for enhanced multimodal conversational emotion recognition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4055–4065, Abu Dhabi, UAE. Association for Computational Linguistics.

Huajian Xin, Z. Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Qihao Zhu, Dejian Yang, Zhibin Gou, Z. F. Wu, Fuli Luo, and Chong Ruan. 2024. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search.

Qi Yan, Zheng Jiang, Zachary Harbin, Preston H Tolbert, and Mark G Davies. 2021. Exploring the relationship between electronic health records and provider burnout: A systematic review. *Journal of the American Medical Informatics Association*, 28(5):1009–1021.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, 14(4):3269—3280.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California. Association for Computational Linguistics.

林巍. 2022. 多任務維度型情感分析之研究. 博士論文, 元智大學, 臺灣博碩士論文知識加值系統.

Hey Vergil at ROCLING-2025 Shared Task: Emotion-Space-Based System for Doctors' Self-Reflection Sentiment Analysis

Ting-Yi Lin *Cong-Ying Lin* *Jui-Feng Yeh*

Department of Computer Science and Information Engineering
National Chiayi University
Chiayi, Taiwan

thomas60611@gmail.com transformers0514@gmail.com ralph@mail.ncyu.edu.tw

摘要

本研究針對 **ROCLING 2025** 維度情感分析任務，提出 **EmoTracer**，一個基於情緒空間的醫師日誌情感分析系統。系統採用 XLNet、BERT 與 LSTM 模型，並以 **SLAKE 病症資料集**及中文資料集（如 **Chinese EmoBank**、**NRC-VAD**）訓練，以捕捉醫師在撰寫病症相關日誌時可能產生的情緒波動。EmoTracer 可將文本轉換為 **Valence** 與 **Arousal** 分數，實驗結果顯示準確率約 60%，皮爾森相關係數（PCC）達 0.9，均方誤差（MAE）約 0.3，可作為心理健康管理的參考工具。系統同時建立了簡易的 **前端 UI**，方便使用者輸入文本並查看分析結果，以完整呈現 **EmoTracer 系統**功能。

Abstract

In the ROCLING 2025 dimensional sentiment analysis task, we present EmoTracer. It is an emotion-space-based system for analyzing doctors' self-reflection texts. The system uses XLNet, BERT, and LSTM models. It is trained on the SLAKE medical dataset and Chinese datasets, such as Chinese EmoBank and NRC-VAD. This helps the system capture the possible emotional changes of doctors when they write patient-related reflections. EmoTracer converts texts into Valence and Arousal scores. The experiments show about 60% accuracy, a Pearson correlation coefficient (PCC) of 0.9, and a mean absolute error (MAE) of 0.3. These results can help support mental health management. The system also has a simple front-end UI. Users can enter texts and see the analysis results. This demonstrates the full functionality of the EmoTracer system.

關鍵字：情緒空間座標、文本情感分析、醫師自我反思文本

Keywords: Emotion Space Coordinates, Text Sentiment Analysis, Doctors' Self-Reflection Texts

1 介紹

近年來，醫師在臨床工作中面臨高度壓力，心理健康問題日益受到關注。過勞、患者照護壓力以及醫療決策責任都可能導致醫師產生焦慮、情緒波動，甚至心理危機。研究顯示，醫師是自殺的高風險族群，其標準化死亡比（SMR）為 1.44，女性醫師的自殺風險更高（SMR = 1.9），顯示這個族群的心理健康問題不容忽視。由於醫師往往忙於工作，缺乏充分的心理支持，早期察覺心理困擾變得困難。

本論文提出 **EmoTracer**，一個基於情緒空間的醫師自我反思日誌分析系統。醫師可透過**表達性書寫（Expressive Writing）**記錄日常臨床經驗與情緒感受，系統則利用 **自然語言處理（NLP）** 將日誌文本轉換為 **Valence** 與 **Arousal** 的二維情緒座標。EmoTracer 可幫助醫師自我識別情緒波動、調整心理狀態，並作為專業心理輔導的輔助工具。

透過對自我反思文本的持續追蹤與分析，醫師的情緒歷程可視化呈現，不僅增進個人自我認知，也能協助醫療機構掌握團隊壓力情況，為心理健康干預提供有效數據支援。

2 模型架構

本論文提出的模型採用了聚合模型 (Aggregation Model) 的方式，融合了 XLNet、

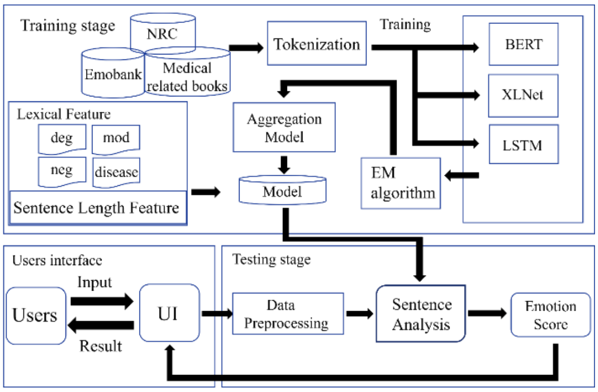
Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 407–412
November 20-22, 2025 ©2025 Association for Computational Linguistics

BERT 以及 LSTM 三種模型，並使用 最大期望演算法 (EM Algorithm) 來動態分配各模型的權重，以彌補彼此的優缺點，並提高整體預測的準確性 。

在詞彙特徵標註方面，我們根據 Chinese EmoBank 和 SLAKE 醫療資料集，標註了病症詞、程度詞、語氣詞及否定詞等詞彙特徵，以擴充詞庫並提升準確度 。同時也針對雙重否定進行數值調整，並利用語句長短對情緒空間進行線性處理，以更精確地捕捉情緒的強度 。

我們使用獨立的測試集來驗證模型效能，並採用平均絕對誤差 (MAE)、皮爾遜相關係數 (PCC) 及正確率 (ACC) 作為評估指標 。最終， EmoTracer 日誌情感分析系統在正負性 (Valence) 預測上的正確率接近 60%，PCC 值超過 0.9，且 MAE 僅有 0.3 左右，顯示出優異的效能 。

最後結合 Python 視窗使用者介面，讓使用者能夠輸入文字或檔案進行分析，並透過二維情緒空間圖與長期情緒變化追蹤圖表，將量化的情緒指標視覺化，有助於識別學生在壓力下的情緒波動，並為預防心理危機提供依據 。



3 實驗方法

為了驗證系統效能，我們設計了兩個實驗，分別為模型準確率測試及語句特徵比較。

在模型準確率測試中，本論文使用準確率 (ACC)、皮爾遜相關係數 (PCC) 和均方誤差 (MAE) 三種指標來評估模型效果 。我們測試了多種模型，包括使用單一資料集 (Chinese EmoBank 或 NRC-VAD) 訓練的 XLNet、LSTM 和 BERT 模型 。其中，未加入詞彙特徵的 XLNet 模型被設為基準線 (Baseline)。最終，我們將所有模型合併成一個聚合模型 (Aggregation Model)，並測試其效能 。

在語句特徵比較實驗中，我們針對不同長度的語句 (詞數分別為 10、18、20、30) 進行了線性調整，並評估其對聚合模型效能的影響 。

微調方式

本論文的聚合模型根據最大期望演算法 (EM Algorithm) 對模型進行微調。由於同時使用 NRC-VAD 和 Chinese EmoBank 兩種資料集進行訓練，我們將其擴展為六個獨立的子模型 。最大期望演算法能夠動態地為這六個模型分配權重，其目的是整合各模型的優點，使其強弱項互相補足，並透過最小化整體預測誤差的目標函數來持續更新權重，最終目標在於提高整體分析的準確性與穩定性 。

為了驗證此優化策略的成效，並全面性地評估模型在情感強度預測這個迴歸問題 (regression problem) 上的性能表現，我們採用了以下三種互補的評估指標：

平均絕對誤差 (Mean Absolute Error, MAE) 是評估迴歸任務的核心指標之一，用於衡量模型預測值與實際標註值之間的平均誤差 。在本研究中，MAE 不僅是評估模型最終性能的關鍵，其概念也與我們模型微調過程中最小化的目標函數緊密相關。MAE 值越小，代表模型的預測在數值上越精準。

皮爾遜相關係數 (Pearson Correlation Coefficient, PCC) 則從另一個維度衡量模型的表現。相較於 MAE 關注預測的誤差大小，PCC 則是用於衡量模型預測結果與實際資料之間的線性相關性 。當 PCC 值越接近 1，表示模型的預測趨勢與真實值的變化趨勢高度一致，證明模型能準確捕捉情感分數的相對高低變化 。

最後，我們引入正確率 (Accuracy, ACC) 作為輔助指標，用以評估模型在判斷情感基本傾向上的表現。在計算上，我們以 Chinese EmoBank 資料集中的 Chinese valence-arousal sentences (CVAS) 子資料集所提供的標註分數作為評估的參考基準。由於情緒標註並不存在絕對的標準答案，因此 ACC 在此處的作用是量化模型預測結果與此參考基準的相符程度。提供了一個具實用價值的互補視角，用以驗證模型在捕捉情感方向的可靠性，從而與 MAE 和 PCC 共同構成了對模型更完整的性能驗證。

資料集

在訓練資料方面，我們採用了三個資料集(如表 1)，其中 Chinese EmoBank： 包含 11,043 筆數據，涵蓋單字、短語、單句和多句文本，並標註了情感的正負性與喚醒值 。而 NRC-VAD： 包含 119,791 筆中文資料，同樣採用情緒空間理論進行標註 。還有醫學生日誌相關的其他書籍，用於擴充領域詞彙 。

詞彙特徵標註的部分，我們根據 Chinese EmoBank 和 SLAKE 醫療資料集，標註了病症詞、程度詞、語氣詞和否定詞，以增強模型對醫療語境的識別能力 。

我們的測試資料集採用訓練資料集 Chinese EmoBank 中的 Chinese valence-arousal sentences (CVAS)，以該資料集的短句標註分數做為參考標準 。DSA-MST 比賽測試資料集： 一個獨立於訓練資料的測試集，用於評估模型的泛化能力 。

詞彙特徵處理

在情感分析系統中，詞彙特徵處理扮演著至關重要的角色。它不僅讓模型能夠識別語言中的情感，更能精確捕捉情感的強度、語氣與方向。我們將這個過程分為以下幾點：

1. 病症詞標註(2216 筆)：

本論文參考了醫療領域的 SLAKE 資料集，對日誌中常見的疾病詞彙（如「感冒」、「頭痛」、「失眠」等）進行了情感值標註。這樣做的目的是為了讓模型能夠理解，這些詞彙本身就帶有負面的情感和較高的喚醒值，從而更精準地識別醫學日誌中因身體狀況而產生的情緒。

2. 程度詞與語氣詞(共 65 筆)：

程度詞（如「非常」、「略微」）與語氣詞（如「也許」、「一定」）能夠為情感提供細緻的層次感。我們根據 Chinese EmoBank 的標準差和模型測試結果，為程度詞分配了不同的調整倍率，讓模型能夠區分「有點不開心」和「超級不開心」之間的強度差異。同樣地，語氣詞和標點符號的結合（如「?」和「!」）則能幫助模型識別情感中的不確定性、推測或強烈的情緒表達。

3. 否定詞與雙重否定(13 筆)：

否定詞（如「不」、「沒」）不僅能反轉情感的正負性，雙重否定（如「不得不」）更能強化語氣。我們針對否定詞進行了數值調整，並特別處理了雙重否定的情況，以確保模型能夠正確理解語義的轉變，例如將一個正向情感的表達轉為強烈的負面情緒。

透過這些細緻的詞彙特徵處理，本論文的聚合模型能夠超越傳統的情感分類，深入理解複雜的情緒變化，讓日誌情感分析系統的結果更加細膩且準確。

語句特徵比較

根據語言簡潔性對情緒空間的影響，中文語句在表達情感時能夠迅速、強烈地傳達情感，尤其是在簡短的語句中，情緒通常會顯得更加急迫、緊張或命令式。例如，簡單的祈使句如「去做功課」、「快走！」或「別說話」都能立即將情感強烈地傳達出來，且這些語句通常會省略主語，使得情緒表達更加直接。

在語句特徵比較的演算法中，我們通過對不同長度的語句進行分析，對語句的詞數進行篩選，並根據語句的詞數分別為 10、18、20、30 來計算斜率及截距。這樣的分析有助於更精確地捕捉情緒的強度，尤其是短語句所表達的強烈情感，並能在情緒空間中準確地映射出其對應的情感強度。

表 1：資料集數量

資料集種類	資料使用數量
NRC-VAD	119791
Chinese EmoBank	11043
其他書籍	92973
程度詞類	223
疾病詞 SLAKE	2216

4 實驗結果

綜合表 2 的所有實驗數據與分析，我們發現本論文提出的聚合模型(Aggregation Model)在情感分析的表現上取得了顯著的成功。相較於單一模型，本論文的聚合模型在所有測試中展現了更為優異且穩定的效能。

在單一模型中，使用 EmoBank 資料集訓練的 BERT 模型在正負性（Valence）的準確性和相關性上表現突出，正確率高達 58.32%，皮爾遜相關係數(PCC)達到 0.8701。然而，

BERT 在喚醒度（Arousal）的預測上誤差較大，而 XLNet 與 LSTM 則在這方面有更好的表現。

透過 最大期望演算法 (EM Algorithm) 的微調與權重分配，我們的聚合模型有效地融合了各個單一模型的優點，並顯著彌補了它們的不足。實驗結果顯示，聚合模型在正負性（Valence）與喚醒度（Arousal）的 正確率(ACC)都超過 50%，分別達到了 59.02% 和 50.97%。

另外，該模型在正負性(Valence)的皮爾遜相關係數(PCC)為 0.9120，已接近上限值，同時其均方誤差（MAE）僅為 0.3146，是所有實驗組別中最低的。證明了聚合模型在預測的精準度上達到了最佳表現，能夠為日誌情感分析提供一個可靠且高效的解決方案。

表 3 的語句特徵比較實驗結果顯示，在語句長度接近 18 個詞時，聚合模型在正負性（Valence）與喚醒值（Arousal）上的 正確率（ACC）都有所提升。雖然這項處理使得皮爾遜相關係數（PCC）與均方誤差（MAE）略有下降，但 MAE 仍維持在 0.3 左右，顯示其誤差並未明顯增加。這證實本論文的語句長短特徵處理，確實能對整體模型的效能有所助益。

表 2 對 Aggregation Model 測試比較

模型	資料集	Valence Accuracy	Arousal Accuracy	PCC Valence	PCC Arousal	MAE Valence	MAE Arousal
Baseline XLNet	EmoBank	41.51%	47.06%	0.8112	0.2215	0.6541	1.0772
XLNet	EmoBank	54.33%	47.28%	0.7809	0.5508	0.5461	0.4295
XLNet	NRC	36.37%	48.14%	0.7899	0.0775	0.5962	0.4134
LSTM	EmoBank	55.65%	47.97%	0.5796	0.4885	0.6753	0.5593
LSTM	NRC	31.56%	48.87%	0.8185	0.7868	0.5868	0.4366
BERT	EmoBank	58.32%	50.56%	0.8701	0.5587	0.5172	0.8432
BERT	NRC	55.44%	44.57%	0.8112	0.1749	0.6283	1.1280
Aggregation Model	ALL	59.02%	50.97%	0.9120	0.6283	0.3146	0.3061
比賽結果	ALL			1.01	0.21	0.63	0.62

表 3 語句特徵比較之實驗結果

語句詞數	Valence Accuracy	Arousal Accuracy	PCC Valence	PCC Arousal	MAE Valence	MAE Arousal
未針對語句特徵做處理	59.02%	50.97%	0.9120	0.6283	0.3146	0.3061
10	59.02%	51.03%	0.9096	0.5538	0.3127	0.3911
18	59.13%	51.34%	0.9318	0.5679	0.3130	0.3957
20	58.81%	50.17%	0.9111	0.5860	0.3131	0.4026
30	58.03%	48.46%	0.9193	0.5066	0.3066	0.4488

這個結果與語言的簡潔性理論相符，亦即簡短的語句通常會更直接、更強烈地表達情感。透過分析與調整，我們的系統能更精準地捕捉日誌中因語句長短所帶來的細微情

緒變化，特別是短句所蘊含的強烈情感，進而提升情感分析的準確度。

5 結論

本論文透過自然語言處理技術，開發了一個能夠精準分析情緒日誌並長期追蹤情感變化的系統。我們深入研究了中文的語言特性，特別是針對否定詞、程度詞等詞彙特徵進行處理，有效地提升了模型的可靠性。

實驗結果顯示，本論文所提出的模型能夠將醫生的自我反思文本轉換為易於理解的情緒空間。在正負性（Valence）的預測上，正確率接近 60%，皮爾遜相關係數（PCC）超過 0.9，且均方誤差（MAE）維持在 0.3 左右，證明了系統在情感分析上的優越性能。

本系統不僅為醫生提供了一個能夠記錄與分析日誌內容的工具，也能透過持續追蹤的功能為其心理健康管理提供參考，幫助他們及時了解自身的情感波動，識別出潛在的心理疾病風險，提供早期預警的支持工具。

6 References

Eva S. Schernhammer and Graham A. Colditz. 2004. Suicide rates among physicians: a quantitative and gender-specific review of the literature. American Journal of Psychiatry, 161(12):2295–2302.

J. M. Smyth. 1998. Written emotional expression: effect sizes, outcome types, and moderating variables. Journal of consulting and clinical psychology, 66(1):174.

J. W. Pennebaker and C. K. Chung. 2011. Expressive Writing: Connections to Physical and Mental Health.

Lung-Hao Lee, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. 2025. ROCLING-2025 Shared Task: Chinese Dimensional Sentiment Analysis for Medical Self-Reflection Texts. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in Neural Information Processing Systems, volume 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

J. H. Wang, T. W. Liu, X. Luo, and L. Wang. 2018. An LSTM approach to short text sentiment classification with word embeddings. In Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018), pages 214–223.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22.

Lung-Hao Lee, Jian-Hong Li and Liang-Chih Yu. 2022. Chinese EmoBank: Building Valence-Arousal Resources for Dimensional Sentiment Analysis. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(4): article 65.

Huan-Ling Lin, Yu-Sheng Lu, Jheng-Wei Chen, et al. 2021. SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical VQA. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9506–9513.

Pratibha Chauhan and Nitin Sharma. 2023. A systematic review on dimensional sentiment analysis. Multimedia Tools and Applications, 82(12):18011–18043.

Jing Zhao, Siyu Kang, Peijie Liu, Gerard de Melo, and Yaling Zhang. 2023. VADER-based iterative deep multi-task learning for valence-arousal prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, 37(11):13264–13272.

Zhen Li, Jian-Hua Li, and Shi-Feng Wang. 2024. Context-Aware and Speaker-Sensitive Network for Dimensional Emotion Recognition in Conversations. IEEE Transactions on Affective Computing.

F. Miedema and S. Bhulai. 2018. Sentiment analysis with long short-term memory networks. Vrije Universiteit Amsterdam, pages 1-17.

B. Huang, Y. Ou, and K. M. Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In Proceedings of the 11th International Conference on Social, Cultural, and Behavioral Modeling (SBP-BRiMS 2018), pages 197–206.

W. Wang, B. Bi, M. Yan, C. Wu, Z. Bao, J. Xia, and L. Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. arXiv preprint arXiv:1908.04577.

M. V. Koroteev. 2021. BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

A. F. Adoma, N. M. Henry, and W. Chen. 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pages 117–121.

H. F. Zhang, C. Zeng, and P. He. 2022. An Emotion Cause Detection Method Based on XLNet and Contrastive Learning. In Proceedings of the International Conference on Software Engineering and Knowledge Engineering (SEKE), pages 646–649.

N. Habbat, H. Anoun, and L. Hassouni. 2022. Combination of GRU and CNN deep learning models for sentiment analysis on French customer reviews using XLNet model. IEEE Engineering Management Review, 51(1):41–51.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, et al. 2016. Building Chinese Affective Resources in Valence-Arousal Dimensions. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, pages 540–545.

KOLab at ROCLING-2025 Shared Task: Research on Emotional Dimensions in Chinese Medical Self-Reflection Texts

Chia-Yu Chan, Chia-Wen Wang, Jui-Feng Yeh
Department of Computer Science and Information Engineering, National Chiayi University
{s1112929,s1112946,ralph}@mail.ncyu.edu.tw

摘要

目前大多數的情感分析技術主要應用於一般性文本，如社交媒體或新聞報導，對於醫學領域中的情感識別仍屬相對空白。自我反思包含個人與內在自我的交流，對於人們未來的生活有正向影響。本篇旨在針對醫學領域相關人員的反思文本，設計出回歸模型，以填補醫學領域在情感分析的空缺。本次任務採用 BERT 模型，配合 Chinese EmoBank 的資料集進行訓練，以 ROCLING 2025 Dimensional Sentiment Analysis – Shared Task 所提供的測試集進行評估，評估結果顯示 Valence 和 Arousal 的 PCC 分別是 0.76、0.58；而 MAE 的分數分別為 0.53、0.82。

Abstract

Currently, most sentiment analysis techniques are primarily applied to general texts such as social media or news reports, and there is still a relative gap in emotion recognition within the medical field. Self-reflection involves communication between individuals and their inner selves, which has a positive impact on people's future lives. This article aims to design a classification model for reflective texts aimed at medical professionals to fill gaps in sentiment analysis within the medical field. This task used a BERT model, trained on a dataset from the Chinese EmoBank, and evaluated using the test set provided by the ROCLING 2025 Dimensional Sentiment Analysis – Shared Task. The assessment results show that Valence and Arousal's PCC scores are 0.76 and 0.58 respectively, while the MAE scores are 0.53 and 0.82, respectively.

關鍵字：BERT、皮爾森相關係數、平均絕對誤差

Keywords : BERT, PCC, MAE

1 Introduction

近年來，隨著深度學習技術日漸發達，本研究使用雙向編碼器 BERT (Bidirectional Encoder Representations from Transformers) 實現一個給醫學領域人員記錄並反思自己的文本。由於 BERT 可以針對許多自然語言處理 NLP (Natural Language Processing) 任務進行微調且應用範圍廣泛，微調（fine-tuning）對模型效能有明顯影響，特定領域語料的再訓練（domain-specific retraining）也更能提升準確性[10]。本研究結合了自然語言處裡及分類式情緒分析以紀錄醫學領域人員的情感分數。使用維度式的情緒分析比傳統情緒分類法（如正面／負面）可以提供更細緻的分析[12]。

有別於英文，中文文本在語意上經常面臨許多問題，如分詞、語意等，因此本研究使用 BertTokenizer 進行分詞，採用 WordPiece 方法，適合用於中文 BERT 模型，將文本切分成子詞 (subword)，再透過特徵向量和詞嵌入的方式回歸 Valence 和 Arousal 的分數。

在文本情緒分析的領域中，Calvo 與 Kim（2013）比較了兩種主要的情緒分析模式，以情緒標籤為基礎的「分類式」和以連續情緒維度（Valence、Arousal、Dominance）為基礎的「維度式」。研究顯示，維度模型能在心理學展現優勢。為後續將心理理論整合到自然語言處理的情緒辨識提供了重要基礎[2]。

2 METHODS

2.1 Model Architecture

本研究設計了一個模型來完成本次任務，使用 BERT 中文預訓練模型 (bert-base-chinese) 進

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 413–417
November 20-22, 2025 ©2025 Association for Computational Linguistics

行，並實現分詞。本模型的任務為提取文本中的每個詞的 Valence 和 Arousal 的特徵向量，去進行 Valence 和 Arousal 的分數計算。

2.1.1 BERT

BERT 模型，主要為 NLP 模型中的編碼器。框架包括兩個步驟：預訓練和微調。 在預訓練期間，模型在不同的預訓練任務上對未標記的資料進行訓練。 而在微調方面，BERT 模型首先用預訓練參數進行初始化，然後使用來自下游任務的標記資料微調所有參數。 雖然每個下游任務都有單獨的微調模型，但它們都是用相同的預訓練參數初始化的[4]。

在預訓練中採用 masked language model (MLM) 是 BERT 重要的預訓練任務，用於建立雙向上下文推理能力[3]。其學習目標是預測文本中被隨機遮蔽的詞，利用上下文兩側的信息進行推斷；Next Sentence Prediction (NSP)，為了提升 BERT 模型在捕捉長程依存關係（long-term dependencies）方面的能力，訓練過程中引入了下一句預測任務（Next Sentence Prediction, NSP），在此任務中，模型需判斷「序列 B 是否為序列 A 的後續內容」。若屬於，序列 A 與序列 B 會從同一文件中依照自然順序抽取；若不屬於，則序列 A 與序列 B 會隨機取樣。透過此設計，模型得以學習文本片段之間的語義連貫性與上下文關聯性，進而加強對篇章結構的理解能力[6]。BERT 的雙向 Transformer 架構能捕捉上下文語義，在 NLP 任務中表現突出[9]。

2.1.2 BertTokenizer

分詞（Tokenization）是指將輸入文本劃分為子單位，稱為詞元（tokens）。這些詞元之後會被用於自然語言處理的後續步驟，例如形態分析（morphological analysis）、詞性標註（word-class tagging）以及句法分析（parsing）[5]。

Tokenizer 是 NLP 的核心元件之一。 由於模型只能處理數字，因此需要標記器 Tokenizer 將輸入的文本轉換為模型可處理的數據。

2.1.3 WordPiece

WordPiece 是 Google 為預訓練 BERT 而開發的標記化算法[11]，其在訓練方面與 BERT 相似，但實際標記化方式不同。WordPiece 是從一個小詞彙表開始，包括模型使用的特殊標記和初始字母表。WordPiece 將文本切分成子詞單位，如「學習中文」可能被分成 ['學'， '##習'， '中'， '##文']，且每個單詞最初是通過將該前綴添加到單詞內的所有字符來拆分的，前綴 ## 表示這個 token 是前一個 token 的延伸。

2.1.4 模型架構

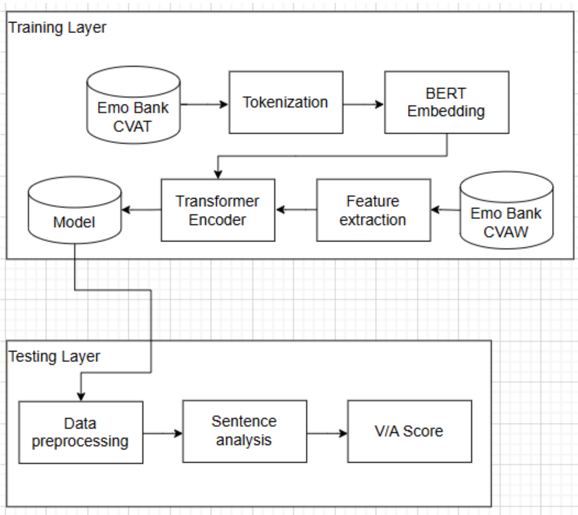


圖 1. 模型架構圖

2.2 Tokenization and Encoding

在本研究中採用 BERT 中文模型 (bert-base-chinese) 內建的 WordPiece 分詞系統對輸入文本進行分詞。模型在輸入文本時首先被轉換為 token 序列，並限制最大長度為 256。並且每個 token 帶有位置的編碼以及注意力遮罩 (attention mask)，以符合 BERT 的輸入格式。

2.3 Feature extraction

在本次任務中使用了 CVAW 作為詞典。當文本輸入並進行分詞後，模型將會檢查文本中的詞彙是否存在詞典中，若存在將提取相對應的 Valence 和 Arousal 的特徵，若不存在將會將數值補充為 [0.0, 0.0]，並生成對應的情緒特徵向量，供模型後續使用。

2.4 Word embedding

詞嵌入（Word Embedding）是一種詞彙表示方法，透過將詞語嵌入到實數向量空間中，將離散的文字轉換為連續數值，並使語意相近的詞語在向量空間中具有相近的位置表示[1]。

在本研究中採用的 BERT 模型中，詞嵌入 (Word Embedding) 可以將文字符號轉為模型可處理的向量表示。在 BERT 中包含了三種嵌入，分別是詞嵌入 (Token Embeddings)、分段嵌入 (Segment Embeddings) 和序列位置的嵌入 (Position Embeddings)。詞嵌入負責將詞彙轉換為向量表示，分段嵌入用於區分不同句子的片段，而位置嵌入則為序列中的每個 token 添加位置訊息。三者相加後形成最終的輸入嵌入向量，進一步輸入模型中的編碼器，完成後續的上下文語意建模。

為了更好的得到文本中的情緒特徵，本研究在 BERT 原始的嵌入基礎上額外使用了 CVAW 中文情緒字典的情緒向量（Valence 與 Arousal 分數），並將其與 BERT 的嵌入層進行融合。使得模型在獲取上下文語意的同時，輸入來自 CVAW 中文字的情緒特徵，為每個 Token 提供額外的 Valence 和 Arousal 的資訊。兩種向量在嵌入層融合後再經由線性層進行回歸，最終同時輸出 Valence 和 Arousal 的預測值。

2.5 Multi task

本研究將 Valence 與 Arousal 的預測視為一個多任務迴歸問題。對於輸入的文本，模型會同時輸出 Valence 預測值與 Arousal 預測值。故採用了多任務加權去強化模型在學習計算 Valence 跟 Arousal 的權重，並使用均方誤差 (MSE) 作為基礎去計算損失函數。其中均方誤差 MSE 和損失函數 (Loss) 的計算如下列公式所示。

$$\text{MSE} = \frac{1}{n} * \sum_{i=1}^{n} (y_i - y'_i)^2$$

$$\text{Loss} = \alpha * \text{MSE(Valence)} + \beta * \text{MSE(Arousal)}$$

MSE 公式中，其中 n 代表樣本數量，y_i代表實際分數，y'_i代表模型預測出的分數。在 Loss 公式中，α 代表 Valence 在模型學習的權重，β 則代表 Arousal 的權重，在本研究中所占權重分別為 0.6 和 0.4，此設計使模型在訓練過程中能更側重於效價的準確預測，同時兼顧喚醒度的表現，達到兩者間的平衡。

3 EXPERIMENTS

3.1 Datasets

本次任務所使用的資料集為中文情緒資料集 Chinese Emobank [7] 中的中文情緒文本 (CVAT) 及中文情緒字典 (CVAW)。其中的資料包含中文單字 (或文本) 並包含了情緒效價 (Valence) 和喚醒度 (Arousal) 的分數，均以浮點數標註。情緒效價表示情緒的正面和負面情緒的程度，喚醒度表示平靜和興奮的程度。兩個維度的數值範圍均為 1（非常消極或平靜）到 9（非常積極或興奮）。

中文情緒字典 (CVAW) 共包含了 5,512 個單字，中文情緒文本 (CVAT) 共包含了 2,969 句中文文本。

3.2 Authentication and Evaluation

本研究中使用了皮爾森相關係數 (PCC) 和平均絕對誤差 (MAE) 作為評估模型的指標。其中 PCC 是一個介於 -1 和 1 之間的值，用於衡量實際值和預測值之間的線性相關性；而 MAE 的理論範圍最低值為 0、最高值為 8。而越低的 MAE 值和越高的 PCC 值，代表了模型預測效能準確越高。兩者指標定義如下列公式所示。

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |a_i - p_i|$$

$$\text{PCC} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{a_i - \mu_A}{\sigma_A}\right)\left(\frac{p_i - \mu_P}{\sigma_P}\right)$$

在 MAE 的公式中，其中 n 代表樣本數量，a_i 為預測值，p_i是真實值。在 PCC 的公式中，其中 $\left(\frac{a_i - \mu_A}{\sigma_A}\right)$、$\mu_A$ 及 σ_A 分別是 a_i 的樣本的標準分數、樣本平均值和樣本標準差。

4 RESULTS

4.1 Validation dataset

在本次任務中，此模型在驗證集的表現如下表所示。評估結果顯示 Valence 和 Arousal 的

PCC 分別是 0.716、0.508；而 MAE 的分數分別為 0.613、1.079。

Valence PCC	Valence MAE	Arousal PCC	Arousal MAE
0.716	0.613	0.508	1.079

表 1. 驗證集中的表現

4.2 Test dataset

在本次任務中，此模型的在測試集實驗結果如下表所示，評估結果顯示 Valence 和 Arousal 的 PCC 分別是 0.76、0.58；而 MAE 的分數分別為 0.53、0.82。在本次競賽[8]中的最終排名得到了第五名的成果。

Valence PCC	Valence MAE	Arousal PCC	Arousal MAE
0.76	0.53	0.58	0.82

表 2. 實驗結果

5 CONCLUSIONS

在本次的任務中，主要運用了 BERT 中文預訓練模型，其中用了 BertTokenizer 進行分詞，，從文本的詞彙中提取特徵向量並融合透過詞嵌入得到的文字向量，從而得到 MAE 及 PCC 的數值。此外，此研究也透過多任務加權的方式來調整模型在 Valence 和 Arousal 分數上學習的權重讓模型可以更好的針對不足的方面加強。在實驗過程中發現模型仍有需加強，使其有更好的結果。未來將運用於醫療領域，可更加清楚的得知醫療相關人員內心真實想法，也可作為醫護人員壓力偵測與心理輔助的基礎模型。

References

Agrawal, Ankur N. 2021. *Introduction to word embeddings. In Hands-on Question Answering Systems with BERT.* Berkeley, CA: Apress.

Calvo, R. A., & Kim, S. M. 2013. *Emotions in text: Dimensional and categorical models. Computational Intelligence*, 29(3), 527–543.

Cui, Yiming, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. *Pre-training with whole word masking for Chinese BERT.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3504–3514.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT 2019), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Grefenstette, Gregory. 1999. *Tokenization. In Syntactic Wordclass Tagging,* pages 117–133. Dordrecht: Springer Netherlands.

Koroteev, M. V. 2021. *BERT: A review of applications in natural language processing and understanding*. arXiv preprint arXiv:2103.11943.

Lee, Lung-Hao, Jian-Hong Li, and Liang-Chih Yu. 2022. *Chinese EmoBank: Building valence–arousal resources for dimensional sentiment analysis. ACM Transactions on Asian and Low-Resource Language Information Processing,* 21(4):Article 65.

Lee, Lung-Hao, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. *2025. ROCLING-2025 Shared Task: Chinese dimensional sentiment analysis for medical self-reflection texts. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025).*

Rietzler, Alexander, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. *Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification.* arXiv preprint arXiv:1908.11860.

Wu, Yichao, Zhengyu Jin, Chenxi Shi, Penghao Liang, and Tong Zhan. 2024. *Research on the application of deep learning-based BERT model in sentiment analysis.* arXiv preprint arXiv:2403.08217.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, ... and Jeff Dean. 2016. *Google's neural machine translation system: Bridging the gap between human and machine translation.* arXiv preprint arXiv:1609.08144.

Yu, Liang-Chih, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. *Building Chinese affective resources in valence–arousal dimensions.*

SCUNLP at ROCLING-2025 Shared Task: Systematic Guideline Refinement for Continuous Value Prediction with Outlier-Driven LLM Feedback

Hong-Rui Pan, Jheng-Long Wu
Department of Data Science
Soochow University
Taiwan, Taipei
Contact: rui0000525@gmail.com

Abstract

Regression-based prediction is widely applied to continuous outputs, such as emotion dimension estimation. However, traditional methods struggle to handle unclear annotation standards and ambiguous cases. To address this challenge, we propose a dual-layer agent-executor framework, where the agent is responsible for constructing and refining guidelines, while the executor applies these guidelines to annotate large-scale data. Notably, we introduce a novel refinement mechanism that can detect outlier instances and provide feedback to the agent for guideline revision, thereby achieving iterative improvement. We applied this method to the ROCLING 2025 shared task (Lee et al., 2025) for predicting valence-arousal (VA) values in medical self-reflection texts. Compared to the unmodified version, the outlier-driven configuration effectively reduced MAE for both V/A, with A-MAE significantly decreased by 7.7%. The final valence-MAE was 0.51 and arousal-MAE was 0.87, ranking fourth.

Keywords: LLM Prediction, Dimensional Sentiment Analysis, Prompt Optimization

1 Introduction

Dimensional emotion analysis have highlighted the importance of continuous valence-arousal (VA) prediction for understanding emotional states in text (Russell, 1980; Buechel & Hahn, 2017). These models have demonstrated remarkable capabilities in capturing the nuanced nature of human emotions across various domains, from social media analysis to clinical applications (Mohammad, 2018; Park et al., 2021; Mitsios et al., 2024). However, despite their impressive performance, current approaches struggle with fundamental challenges in specialized domains such as medical self-reflection texts, where emotional expressions are often subtle, context-dependent, and require domain-specific understanding (Teodorescu et al., 2023; Alvarez-Gonzalez et al., 2021). This highlights the need for more adaptive and interpretable frameworks that can systematically improve annotation quality through iterative refinement.

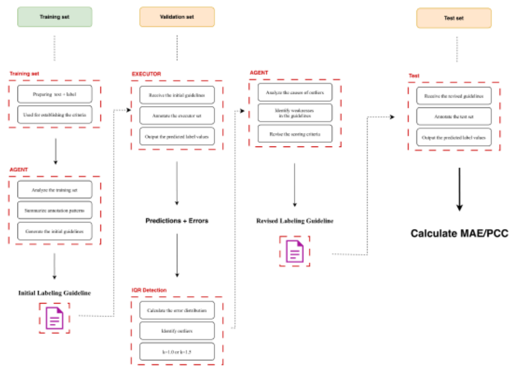


Figure 1: Overview of the dual-layer Agent-Executor framework. The Agent constructs and revises guidelines, while the Executor performs annotations under three configurations, with IQR-based outlier detection driving the feedback loop.

Many current methods for emotion prediction rely on static annotation guidelines and traditional regression approaches that treat valence and arousal dimensions independently (Park et al., 2021; Bobicev & Sokolova, 2018). These approaches typically depend on large-scale manually annotated datasets with consistent labeling criteria. However, there are two main challenges with these methods. Firstly, creating high-quality annotations for dimensional emotion analysis is expensive and time-consuming, particularly in specialized domains like healthcare where expert knowledge is required (Wei et al., 2021; Giachelle et al., 2021). For instance, Wei et al. (2021) demonstrated that "manual annotation

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 418–426
November 20-22, 2025

by clinical experts is both time consuming and expensive," while Giachelle et al. (2021) noted that "manual annotation of large datasets is an expensive and time-consuming task requiring plenty of expert annotators with extensive experience in biomedical contents." Secondly, static guidelines cannot adapt to the diverse and ambiguous emotional expressions encountered in real-world texts, limiting their effectiveness for edge cases and domain-specific nuances (Alvarez-Gonzalez et al., 2021).

Large Language Models (LLMs), on the other hand, have shown remarkable ability to understand complex instructions and provide detailed feedback, indicating potential for more sophisticated annotation and refinement methods (Brown et al., 2020; Ouyang et al., 2022). Recent work in prompt optimization has demonstrated that iterative refinement can significantly improve model performance through systematic feedback incorporation (Wan et al., 2023). Madaan et al. (2023) showed that LLMs can iteratively improve their outputs through self-generated feedback, achieving approximately 20% improvement across various tasks without additional training.

Considering the advantages and disadvantages mentioned above, We propose a dual-layer Agent-Executor framework (Figure 1) for iterative guideline refinement, addressing annotation quality challenges in limited-data settings. In this approach, a high-capacity Agent formulates guidelines by analyzing domain knowledge and annotation complexities, while an efficient Executor applies these guidelines at scale to annotate data. A key feature is an outlier-driven feedback loop that decrease deviations in Executor predictions and feeds them back to the Agent for guideline revision. Evaluated on the ROCLING 2025 shared task for predicting valence-arousal values in Chinese medical self-reflection texts, our framework achieved fourth place. These findings highlight the effectiveness of adaptive guideline refinement and outlier feedback in enhancing annotation consistency and performance in specialized medical emotion prediction tasks.

The main contributions of this work are: (1) A novel dual-layer Agent-Executor framework that separates high-level guideline construction from efficient large-scale annotation; (2) An outlier-driven feedback mechanism that enables systematic identification and correction of problematic predictions; (3) Empirical validation on medical self-reflection texts showing the effectiveness of iterative refinement for dimensional emotion analysis.

2 Related Work

Recent advances in dimensional emotion analysis, prompt optimization, and agent-based NLP frameworks have converged to address the challenges of continuous emotion prediction in specialized domains (Buechel & Hahn, 2017; Madaan et al., 2023; Zhao et al., 2024). In this section, we review three key research areas that inform our approach: dimensional emotion analysis for valence-arousal prediction, iterative prompt refinement methods, and hierarchical agent frameworks for NLP tasks.

2.1 Dimensional Emotion Analysis

Dimensional models of emotion, particularly the valence-arousal framework, have emerged as robust representations for capturing the continuous nature of emotional states in text (Russell, 1980). Buechel and Hahn (2017) established EmoBank, a foundational corpus of 10,000 sentences annotated with Valence-Arousal-Dominance dimensions, demonstrating the superiority of reader's perspective over writer's perspective in terms of inter-annotator agreement. This bi-perspectival approach highlighted the inherent challenges in dimensional emotion annotation, where different viewpoints can lead to substantially different emotional interpretations.

Recent advances have addressed the gap between categorical and dimensional emotion representations. Park et al. (2021) presented a novel approach for predicting fine-grained VAD dimensions from categorical emotion annotations using Earth Mover's Distance loss, showing that traditional regression approaches treating dimensions independently suffer from significant limitations. Mitsios et al. (2024) further advanced the field by introducing ordinal classification techniques for two-dimensional emotion spaces, addressing perceptual similarities among emotional classes and achieving substantial improvements in prediction accuracy.

However, significant challenges persist in dimensional emotion analysis. Bagdon et al. (2024) noted that "humans perform worse when tasked to choose values from a rating scale," highlighting fundamental annotation reliability issues that affect model training. These challenges are compounded in specialized domains such as medical texts, where emotional expressions are often subtle and context-dependent.

2.2 Prompt optimization

The limitations of static prompting approaches have driven significant research into adaptive and iterative prompt optimization methods. Ye et al. (2024) introduced the PE2 framework, which addresses static prompt limitations through meta-prompt components that enable iterative refinement and targeted prompt editing. This approach demonstrated the ability to rectify erroneous prompts and adapt to domain-specific requirements through systematic feedback incorporation.

Self-adaptive prompting has emerged as a key paradigm for dynamic prompt optimization. Wan et al. (2023a) proposed Consistency-based Self-adaptive Prompting (COSP), which dynamically selects examples based on consistency measures, achieving 15% improvement over static baselines. Their subsequent work (Wan et al., 2023b) extended this approach to Universal Self-Adaptive Prompting, automatically selecting suitable queries and responses as pseudo-demonstrations across diverse task types.

Madaan et al. (2023) introduced Self-Refine, demonstrating that large language models can iteratively improve their outputs through self-generated feedback without additional training. Their approach achieved approximately 20% absolute improvement on average across various tasks, establishing the viability of iterative refinement for quality enhancement. This work is particularly relevant to our outlier-driven approach, as it shows how models can identify and correct problematic aspects of their outputs through systematic feedback loops.

2.3 Hierarchical Frameworks

Hierarchical and multi-agent frameworks have shown substantial promise for complex NLP tasks requiring coordinated reasoning and execution. Zhao et al. (2024) presented EPO, a hierarchical LLM agent framework with separate components for subgoal prediction and action generation, achieving first place on the ALFRED leaderboard through effective dual-layer architecture design. This work demonstrates the power of role specialization in agent frameworks.

Wang et al. (2024) explored executable code actions in agent frameworks, showing that dual-component architectures with structured agent-executor separation can achieve 20% higher success rates than monolithic approaches. Their work highlights the importance of clear separation between high-level reasoning and low-level execution components.

Recent work has also addressed the specific challenges of outlier detection and iterative improvement in NLP systems. Zhang et al. (2024) further advanced this area by decomposing LLM confidence into uncertainty and fidelity components, providing the foundation for systematic identification of problematic examples. Hu et al. (2024) demonstrated Self-Refinement Tuning using model-generated feedback for iterative improvement, showing how outlier-driven learning can enhance model performance through systematic identification and correction of problematic outputs.

Our work builds upon these foundations by combining dimensional emotion analysis challenges with iterative prompt refinement techniques within a specialized agent-executor framework, specifically designed for prediction tasks where both accuracy and interpretability are crucial.

3 Method

This section presents our dual-layer Agent-Executor framework for iterative guideline refinement in valence-arousal prediction. We first introduce the overall framework architecture, then detail the Agent and Executor components, followed by our outlier-driven feedback mechanism for systematic guideline improvement.

3.1 Guideline Formulation

The Agent component serves as the rule-making authority responsible for understanding the complexities of dimensional emotion analysis and constructing comprehensive annotation guidelines. The Agent's primary functions encompass theoretical knowledge synthesis, systematic guideline construction, and iterative refinement based on feedback analysis.

Theoretical Foundation Integration: The Agent synthesizes established theoretical frameworks from dimensional emotion literature with empirical patterns observed in annotated text data. It incorporates understanding of the circumplex model of affect while adapting to the nuanced emotional expressions characteristic of specialized text domains.

Initial Guideline Construction Process The Agent employs an open-ended instruction framework designed to enable flexible and comprehensive guideline development. Rather than imposing rigid structural constraints that might limit the model's reasoning capabilities, the initial prompt template encourages creative and thorough guideline formulation, the initial prompt template as shown in Figure 2.

> I will give you a text and its corresponding VA values. Based on this information, please draft a set of guidelines for scoring VA. The guidelines must include separate scoring rubrics (tables) for V and for A. I will use this prompt as the standard for prediction, so please write it carefully.

Figure 2: The template for initial annotation guideline

This unconstrained prompt design allows the Agent to autonomously determine the most appropriate organizational structure and content depth for the annotation guidelines. We found that We found that even without specific formatting requirements, the enables the agent to leverage framework its inherent reasoning capabilities to identify key evaluation dimensions, establish a logical hierarchy, and establish multifaceted criteria. These criteria emerge naturally from the data patterns instead of being restricted by prescriptive predefined rules.

3.2 Annotation Execution

The Executor component is responsible for applying the Agent's guidelines to perform large-scale valence-arousal prediction with rigorous quality control and standardized output formatting.

Once receiving the comprehensive annotation guidelines from the Agent, the executor systematically applies them to the unlabeled text dataset. To ensure high-quality annotations, the component implements strict adherence protocols that require explicit reference to guideline criteria during the prediction process. Each input text undergoes systematic evaluation against the established rubrics, with the Executor required to demonstrate clear reasoning chains linking textual features to scoring decisions.

3.3 Outlier-Driven Feedback Mechanism

The core innovation of our framework lies in the systematic identification and utilization of outlier predictions for guideline improvement. Following initial template-based prediction on the validation set, we employ the Interquartile Range (IQR) method to identify outlier instances based on prediction errors relative to ground truth values.

IQR-Based Outlier Identification We utilize the IQR method for outlier. Unlike methods that rely on standard deviation, IQR provides several key advantages: (1) Distributional Robustness - IQR remains stable even when error distributions are non-normal or contain extreme outliers, making it particularly suitable for emotion prediction tasks where errors may not follow Gaussian distributions; (2) Percentile-Based Thresholds - By defining outliers as values beyond Q1 - k×IQR or Q3 + k×IQR, the method provides interpretable thresholds that correspond to natural data quartiles; (3) Insensitivity to Outliers - Since IQR is calculated using only the 25th and 75th percentiles, it is not influenced by extreme values, preventing the masking effect where true outliers make other outliers appear normal.

For each dimension (valence and arousal), we calculate the absolute prediction error as:

$$Error_i = |y_{pred,i} - y_{true,i}| \tag{1}$$

where $y_{pred,i}$ and $y_{true,i}$ represent the predicted and ground truth values for instance i, respectively. The IQR is then computed as:

$$IQR = Q_3 - Q_1 \tag{2}$$

where Q_1 and Q_3 are the first and third quartiles of the error distribution. Outliers are identified as instances where:

$$Error_i \in (Q_3 + k \times IQR, +\infty) \tag{3}$$

where k is the threshold multiplier. Given that prediction errors are non-negative (absolute values), we primarily focus on the upper bound criterion.

4 Experiment

4.1 Dataset

Our experimental evaluation employs datasets provided by the ROCLING 2025 shared task organizers (Lee et al., 2025).

We utilize The Chinese EmoBank (Lee et al., 2022) for initial guideline construction. Due to our focus on sentence-level emotion recognition and hardware constraints, we extracted 600 instances from the sentence-level (CVAS) and text-level (CVAT) components through stratified sampling to maintain distributional consistency with the complete dataset.

The validation set contains 994 doctors' self-reflection texts for system development, while the test set provides 1,541 doctors' self-reflection texts for final performance evaluation.

4.2 Outlier Detection Configurations

To evaluate the effectiveness of our outlier-driven feedback mechanism, we implement three experimental configurations that systematically assess the impact of iterative guideline refinement:

- **Baseline:** Employing the original prompt template to produce static guidelines that remain unchanged throughout the evaluation process, providing a reference point for measuring improvement.
- **Conservative Detection:** Implementing outlier-driven feedback mechanism with a conservative threshold setting (k = 1.5). Outliers are identified when prediction errors exceed threshold, focusing on the most significant prediction failures to drive targeted guideline improvements while maintaining stability in the refinement process.
- **Aggressive Detection:** This configuration employs a more aggressive threshold setting (k = 1.0) to identify a broader range of prediction anomalies. By lowering the outlier detection threshold to $Q_3 + 1.0 \times IQR$, this approach captures more instances for feedback analysis, enabling comprehensive guideline refinement at the potential cost of including less critical prediction errors.

4.3 Evaluation Metrics

Following the official evaluation protocol established by the ROCLING 2025 shared task organizers, we employ two primary metrics for assessing dimensional emotion prediction performance:

Mean Absolute Error (MAE): The MAE measures the average magnitude of prediction errors without considering their direction:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_{pred,i} - y_{true,i}| \tag{4}$$

where n represents the total number of instances. MAE provides several advantages for dimensional emotion analysis: (1) Interpretability - MAE values directly correspond to the average prediction error in the original scale, making results easily interpretable; (2) Robustness to Outliers - Unlike squared error metrics, MAE is less sensitive to extreme prediction errors, providing a more stable assessment of typical model performance; (3) Linear Penalty - MAE assigns equal weight to all errors regardless of magnitude, avoiding the disproportionate influence of large errors that can skew evaluation in squared metrics.

Pearson Correlation Coefficient (PCC): The PCC measures the linear correlation between predicted and true values:

$$PCC = \frac{\sum_{i=1}^{n}(y_{pred,i} - \bar{y}_{pred})(y_{true,i} - \bar{y}_{true})}{\sqrt{\sum_{i=1}^{n}(y_{pred,i} - \bar{y}_{pred})^2 \sum_{i=1}^{n}(y_{true,i} - \bar{y}_{true})^2}} \tag{5}$$

where ȳ represents the mean values. PCC offers complementary evaluation benefits: (1) Scale Invariance - PCC is unaffected by linear transformations, focusing on the relationship structure rather than absolute values; (2) Ranking Preservation - High PCC values indicate that the model maintains the relative ordering of emotional intensities across instances; (3) Distributional Alignment - PCC captures how well the predicted distribution matches the true distribution pattern, essential for dimensional emotion modeling.

The combination of MAE and PCC provides comprehensive evaluation coverage, with MAE assessing absolute prediction accuracy and PCC evaluating the preservation of emotional intensity relationships across the continuous valence-arousal space.

4.4 Implementation Details

Our dual-layer Agent-Executor framework leverages different model configurations optimized for their respective roles in the annotation pipeline.

Agent: The agent employs GPT-o3 as the underlying language model. This high-capacity model provides the advanced reasoning capabilities necessary for analyzing complex annotation patterns, identifying systematic errors in outlier feedback, and formulating comprehensive guideline refinements that address domain-specific challenges in medical text emotion analysis.

Executor: The Executor utilizes GPT-4o-mini as the base model. This configuration balances annotation quality with computational efficiency, enabling cost-effective processing of the extensive validation and test datasets while maintaining consistent application of the Agent's refined guidelines across all instances.

5 Results

Config	V-MAE	V-PCC	A-MAE	A-PCC
Baseline	0.5095	**0.7676**	0.9381	0.5745
Conservative	**0.5105**	0.7625	**0.8661**	**0.5860**
Aggresive	0.5470	0.5105	0.9964	0.5461

Table 1:Results under different configurations. These configurations all use the same training, validation, and test data sizes.validation, and test data sizes.

Table 1 presents the comparative performance of our dual-layer Agent-Executor framework across three experimental configurations on the ROCLING 2025 shared task. The results demonstrate the effectiveness of our outlier-driven feedback mechanism for improving dimensional emotion prediction in medical self-reflection texts.

Arousal Prediction Benefits More from Iterative Refinement Conservative configuration achieves the best overall valence prediction results, with V-MAE of 0.5105 and V-PCC of 0.7625, representing improvements over the Baseline While the MAE shows marginal improvement, the slight decrease in PCC suggests that the conservative outlier detection may not capture sufficient feedback for substantial correlation enhancement. The Aggressive configuration shows degraded performance, indicating that overly broad outlier detection may introduce noise that compromises guideline quality.

The outlier-driven feedback mechanism demonstrates more pronounced improvements in arousal prediction. The Conservative configuration achieves substantial improvements with A-MAE of 0.8661and A-PCC of 0.5860. This represents a reduction in prediction error by approximately 7.7% and an improvement in correlation by 2.0%. The Aggressive configuration shows mixed results with A-MAE of 0.9964 (worse than baseline) but maintains comparable correlation performance

Aggressive Feedback Threshold Leads to Performance Degradation Conservative approach consistently outperforms both Baseline and Aggressive configurations across most metrics, particularly for arousal prediction. This suggests that targeted identification of the most significant prediction failures provides optimal feedback for guideline refinement without introducing excessive noise. The Aggressive approach appears to suffer from over-correction, where the inclusion of marginal outliers leads to guideline instability and reduced prediction accuracy.

Context Complexity Makes Arousal Assessment More Challenging Than Valence Results reveal interesting asymmetries between valence and arousal prediction improvements. The outlier-driven mechanism shows greater effectiveness for arousal prediction, possibly indicating that arousal-related annotation guidelines benefit more from iterative refinement compared to valence guidelines. This may reflect the inherent complexity of arousal assessment in medical contexts, where emotional intensity can be more ambiguous than emotional polarity.

Overall The results validate our hypothesis that systematic outlier identification and feedback can enhance dimensional emotion prediction, with the Conservative configuration representing the optimal balance between comprehensive feedback and guideline stability.

6 Conclusion

This paper presented a novel dual-layer Agent-Executor framework for iterative guideline refinement in dimensional emotion analysis, specifically designed to address the challenges of valence-arousal prediction in medical self-reflection texts. Our approach systematically combines high-level reasoning capabilities with efficient execution through a hierarchical architecture that enables cost-effective scaling while maintaining annotation quality.

The key innovation lies in our outlier-driven feedback mechanism, which transforms prediction errors from isolated failures into systematic learning opportunities. By employing IQR-based outlier detection, we identified problematic predictions and fed them back to the Agent component for targeted guideline improvements.

This iterative refinement process enables continuous adaptation to domain-specific challenges without requiring extensive manual annotation efforts.

Our experimental evaluation on the ROCLING 2025 shared task demonstrated the effectiveness of this approach. The Conservative outlier detection configuration achieved optimal performance balance, with particularly notable improvements in arousal prediction. The results reveal important insights about dimensional emotion analysis in medical contexts: arousal assessment benefits more from iterative refinement than valence prediction, suggesting that emotional intensity evaluation presents greater annotation challenges than emotional polarity in healthcare narratives.

The framework's practical contributions extend beyond performance improvements. The dual-layer architecture provides a cost-effective solution that leverages expensive high-capacity models only for guideline construction while using efficient models for large-scale annotation. The outlier-driven feedback mechanism offers interpretability through explicit identification of systematic weaknesses, enabling targeted improvements rather than global parameter adjustments.

Limitations and Future Work While our approach shows promising results, several limitations warrant acknowledgment. The framework's effectiveness depends on the quality of initial guidelines, and extremely poor starting points may require multiple refinement iterations. The IQR-based outlier detection, while robust, may not capture all forms of systematic errors, particularly those involving subtle contextual nuances. Future work should explore more sophisticated outlier detection methods that incorporate semantic similarity and domain-specific error patterns.

Additionally, our evaluation focused on a single domain (medical self-reflection texts) and language (Chinese). Extending the framework to multilingual settings and diverse text domains would strengthen its generalizability claims. Investigation of different Agent-Executor model combinations and the integration of human-in-the-loop refinement mechanisms represent promising research directions.

The proposed dual-layer Agent-Executor framework with outlier-driven feedback provides a principled approach to iterative guideline improvement in dimensional emotion analysis, offering both practical benefits for annotation quality and theoretical insights into systematic error correction in specialized domains.

References

Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenç Gómez. 2021. Uncovering the Limits of Text-based Emotion Detection. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2560–2583, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. “You are an expert annotator”: Automatic Best–Worst-Scaling Annotations for Emotion Intensity Modeling. 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 7924–7936, Mexico City, Mexico. Association for Computational Linguistics.

Victoria Bobicev and Marina Sokolova. 2018. Thumbs Up and Down: Sentiment Analysis of Medical Online Forums. In Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task, pages 22–26, Brussels, Belgium. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and

Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pages 1877–1901.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Fabio Giachelle, Ornella Irrera, and Gianmaria Silvello. 2021. MedTAG: a portable and customizable annotation tool for biomedical documents. BMC Medical Informatics and Decision Making, 21(1):1–12.

Chi Hu, Yimin Hu, Hang Cao, Tong Xiao, and JingBo Zhu. 2024. Teaching Language Models to Self-Improve by Learning from Language Feedback. In Findings of the Association for Computational Linguistics: ACL 2024, pages 6090–6101, Bangkok, Thailand. Association for Computational Linguistics.

Lung-Hao Lee, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. 2025. ROCLING-2025 Shared Task: Chinese Dimensional Sentiment Analysis for Medical Self-Reflection Texts. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: Building Valence-Arousal Resources for Dimensional Sentiment Analysis. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(4): Article 65, 1-18.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023), pages 46534–46594.

Michail Mitsios, Georgios Vamvoukakis, Georgia Maniati, Nikolaos Ellinas, Georgios Dimitriou, Konstantinos Markopoulos, Panos Kakoulidis, Alexandra Vioni, Myrsini Christidou, Junkwang Oh, Gunu Jho, Inchul Hwang, Georgios Vardaxoglou, Aimilios Chalamandaris, Pirros Tsiakoulis, and Spyros Raptis. 2024. Improved Text Emotion Prediction Using Combined Valence and Arousal Ordinal Classification. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 808–813, Mexico City, Mexico. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022), pages 27730–27744.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional Emotion Detection from Categorical Emotion. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4367–4380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James A. Russell. 1980. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161–1178.

Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Saif Mohammad. 2023. Language and Mental Health: Measures of Emotion Dynamics from Text as Linguistic Biosocial Markers. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 3117–3133, Singapore. Association for Computational Linguistics.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023a. Better Zero-Shot Reasoning with Self-Adaptive Prompting. In Findings of the Association for Computational Linguistics: ACL 2023, pages 3493–3514, Toronto, Canada. Association for Computational Linguistics.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, Hanjun Dai, Julian Martin Eisenschlos, Sercan Arik, and Tomas Pfister. 2023b. Universal self-adaptive prompting. Computing Research Repository, arXiv:2305.14926.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024. Executable code actions elicit better LLM agents. In

Proceedings of the 41st International Conference on Machine Learning, pages 50682–50695.

Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. 2021. Clinical text annotation: What factors are associated with the cost of time? In AMIA Annual Symposium Proceedings, volume 2018, pages 1552–1560.

Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. Prompt Engineering a Prompt Engineer. In Findings of the Association for Computational Linguistics: ACL 2024, pages 355–385, Bangkok, Thailand. Association for Computational Linguistics.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. Calibrating the Confidence of Large Language Models by Eliciting Fidelity. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2959–2979, Miami, Florida, USA. Association for Computational Linguistics.

Qi Zhao, Haotian Fu, Chen Sun, and George Konidaris. 2024. EPO: Hierarchical LLM agents with environment preference optimization. Computing Research Repository, arXiv:2408.16090.

Taiwanese Hakka Across Taiwan Corpus and Formosa Speech Recognition Challenge 2025 – Dapu & Zhao'an Accents

廖元甫 Yuan-Fu Liao
陽明交大 National Yang Ming Chiao Tung University
yfliao@nycu.edu.tw

郭志忠 Chih-Chung Kuo
陽明交大 National Yang Ming Chiao Tung University
cck@nycu.edu.tw

黃昭世 Chao-Shih Huang
陽明交大 National Yang Ming Chiao Tung University
hsf @nycu.edu.tw

藍瑀翔 Yu-Siang Lan
陽明交大 National Yang Ming Chiao Tung University
yslan168@nycu.edu.tw

賴翰俊 Han-Chun Lai
華碩雲端 ASUS Cloud
patrick.lai@asuscloud.com

許文翰 Wen-Han Hsu
長問科技 BRONCI Technology Inc
jeff.hsu@bronci.com.tw

摘要

為了重振瀕危的臺灣客家話，臺灣首個大規模客語語音語料庫 (HAT) 應運而生，該語料庫預計蒐集涵蓋臺灣各地腔調客家話的錄音。本文介紹 HAT 語料庫近兩年擴充的第二部分：大埔腔和詔安腔。此外，為了推廣此新建語料並評估目前最先進客語 ASR 系統的效能，特舉辦 2025 年福爾摩沙語音辨識挑戰賽－第二屆客語 ASR 競賽 (FSR-2025–Hakka ASR II)。共有 16 隊參加兩個賽道—客語語音辨識轉漢字、客語語音辨識轉拼音。最佳結果分別為：漢字—字元錯誤率 7.50%；拼音—音節錯誤率 14.81%。

Abstract

To revive the endangered Hakka language in Taiwan, the first large-scale Hakka speech corpus covering all aspects of Taiwanese Hakka across Taiwan (HAT) was created. This paper introduces the second part of the HAT corpus: the Dapu and Zhao'an accents. Furthermore, to promote this newly constructed corpus and evaluate the performance of the most advanced Hakka ASR system, the 2025 Formosa Speech Recognition Challenge, FSR-2025–Hakka ASR II, was held. Sixteen teams participated on two tracks: speech-to-Hakka-Hanzi and speech-to Hakka-Pinyin. The best results were: Hanzi character error rate (CER) 7.50%; Pinyin syllable error rate (SER) 14.81%.

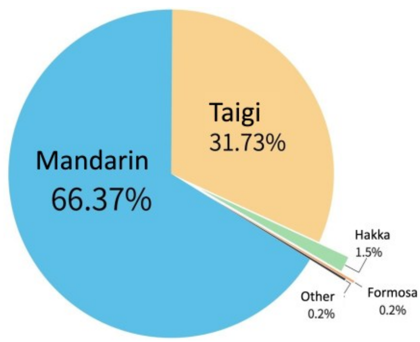


圖 1. 臺灣 2020 年人口及住宅普查報告中，本國籍常住人口主要使用語言統計。

關鍵字：臺灣客語、大埔腔、詔安腔、語音語料庫、自動語音辨識

Keywords: Taiwanese Hakka, Dapu accent, Zhao'an accent, speech corpus, automatic speech recognition (ASR)

1 Introduction

臺灣客語，簡稱為客語，曾是臺灣第二大語言，約有 15-20%的居民使用。然而，由於華語的主導地位，客語如今已瀕臨滅絕。尤其是在 1949 年後，政府積極推廣國語（華語）作為臺灣的官方語言和主要語言。這種推廣導致其他語言的使用率和地位下降，包括臺灣客語。根據臺灣 2020 年人口及住宅普查報告 (主計總處 2020)，目前只有 1.5%的人口使用客語作為主要語言（見圖 1）；以客語為次要使用語言者也僅有 4.0%。換句話說，若不積極復振，客語在臺灣可能很快就會消亡。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 427–434
November 20-22, 2025

圖 2. 臺灣客語五種腔調及其使用地區。

<table>
<tr><td></td><td>調號</td><td colspan="2">1</td><td>5</td><td>2</td><td>3</td><td>7</td><td>4</td><td>8</td></tr>
<tr><td></td><td>調名</td><td colspan="2">陰平</td><td>陽平</td><td>上聲</td><td>陰去</td><td>陽去</td><td>陰入</td><td>陽入</td></tr>
<tr><td rowspan="2">四縣</td><td>調值</td><td colspan="2">24</td><td>11</td><td>31</td><td colspan="2">55</td><td>2</td><td>5</td></tr>
<tr><td>調型</td><td colspan="2">vˊ</td><td>vˇ</td><td>vˋ</td><td colspan="2">v</td><td>vdˋ</td><td>vd</td></tr>
<tr><td rowspan="2">南四縣</td><td>調值</td><td>24</td><td>33</td><td>11</td><td>31</td><td colspan="2">55</td><td>2</td><td>5</td></tr>
<tr><td>調型</td><td>vˊ</td><td>v⁺</td><td>vˇ</td><td>vˋ</td><td colspan="2">v</td><td>vdˋ</td><td>vd</td></tr>
<tr><td rowspan="2">海陸</td><td>調值</td><td colspan="2">53</td><td>55</td><td>24</td><td>11</td><td>33</td><td>5</td><td>2</td></tr>
<tr><td>調型</td><td colspan="2">vˋ</td><td>v</td><td>vˊ</td><td>vˇ</td><td>v+</td><td>vd</td><td>vdˋ</td></tr>
<tr><td rowspan="2">大埔</td><td>調值</td><td>33</td><td>35</td><td>113</td><td>31</td><td colspan="2">53</td><td><u>21</u></td><td><u>54</u></td></tr>
<tr><td>調型</td><td>v⁺</td><td>vˊ</td><td>vˇ</td><td>v^</td><td colspan="2">vˋ</td><td>vdˋ</td><td>vd</td></tr>
<tr><td rowspan="2">饒平</td><td>調值</td><td colspan="2">11</td><td>55</td><td colspan="2">53</td><td>24</td><td>2</td><td>5</td></tr>
<tr><td>調型</td><td colspan="2">vˇ</td><td>v</td><td colspan="2">vˋ</td><td>vˊ</td><td>vdˋ</td><td>vd</td></tr>
<tr><td rowspan="2">詔安</td><td>調值</td><td colspan="2">11</td><td>53</td><td colspan="2">31</td><td>55</td><td><u>24</u></td><td><u>43</u></td></tr>
<tr><td>調型</td><td colspan="2">vˇ</td><td>vˋ</td><td colspan="2">v^</td><td>v</td><td>vdˊ</td><td>vdˋ</td></tr>
</table>

表 1. 六種腔調之臺灣客語拼音聲調表。

要振興客家話，不僅要妥善保存，更要引進一些現代語音人工智慧 (AI) 技術，例如自動語音辨識 (ASR) 和語音合成 (TTS)，以支援人們的日常生活。這些技術將幫助並鼓勵人們使用客家話。因此，一個大型臺灣客語語料庫 HAT (Hakka Across Taiwan)，即「HAT 臺灣客語語音資料庫」應運而生。並於 2023 年推出第一版 HAT-Vol-1 (客委會 2024)，包含臺灣客語最主要的兩大腔調：四縣與海陸。

HAT 語料庫旨在建立自動語音辨識 (ASR) 和語音合成 (TTS) 系統。為此當年也舉辦了 FSR-2023 - Hakka ASR (Formosa Speech Recognition Challenge 2023 - Hakka ASR)，即「福爾摩莎語音辨識挑戰賽 2023—客語自動語音辨識」(Liao et al. 2023；陽明交大 2023)，以推廣 HAT 語料庫，並評估正在開發的先進客家語自動語音辨識系統的性能。FSR-2023 只針對 HAT-Vol-1 所包含的四縣與海陸腔客語。今年隨著其他腔調語料逐漸蒐集完成，特別針對大埔與詔安兩種腔調再次舉辦 FSR-2025 - ASR II 客語語音辨識競賽 (陽明交大 2025)。

本論文將首先簡要說明臺灣客語的腔調種類。其次介紹 HAT 語料庫在 2023 年之後增加的大埔腔和詔安腔，其中包含 251 位說話人，約 400 小時的 ASR 語音數據；以及 4 位說話人，共 125 小時的 TTS 數據。最後，主要說明 FSR-2025 挑戰賽，包括熱身賽與決賽的賽制、使用的資料集、參加的隊伍、比賽成績，以及成績較好的隊伍所使用的技術概述。

2　臺灣客語

臺灣客語是臺灣地區使用的客家話方言，主要由客家人使用。它是臺灣官方認可的國家語言之一，傳統上主要共有五種次方言腔調：四縣、海陸、大埔、饒平、和詔安腔。圖 2 展示了臺灣客家話的五種腔調及其在臺灣的使用區域 (賴維凱 2008)。另外，2011 年教育部國語推行委員會在審查《臺灣客家語常用詞辭典》時，正式決定將南四縣腔從四縣腔分開來；也就是將以屏東內埔為代表的南部六堆客家聚落，與北部以苗栗為代表的四縣腔，其詞彙與腔調變異在辭典中分別出來，稱為南四縣腔。

臺灣五種客家次方言中，使用最廣泛的是四縣和海陸。四縣有六個聲調，起源於廣東梅州；海陸有七個聲調，起源於廣東海豐和陸豐。四縣腔為目前使用人口最多的客家話，幾乎涵蓋北部、中部、南部與東部；海陸腔居次，使用人口主要以新竹、南桃園、苗栗北境及花東一帶。FSR-2023 挑戰賽，就是以四縣與海陸兩種腔調為語音辨識標的。

剩下較少人口的客語腔調中，大埔腔分布在臺中東勢、石岡、新社、和平、豐原及苗栗卓蘭；饒平腔主要分布在桃園、新竹、苗栗，彰化永靖次之；詔安腔是使用人口最少的，以雲林境內的二崙、崙背為主，桃園大溪次之 (賴維凱 2008)。HAT 語料庫目前已經完成大埔、詔安兩種腔調的蒐集，因此今年的 FSR-2025 客語語音辨識挑戰賽就以此兩種語料資源極度稀少的腔調作為競賽標的。

臺灣客語的官方書寫系統是臺灣客語漢字。教育部分別於 2009 年和 2010 年提出了兩批「臺灣客語書寫推薦用字」(教育部 2025/7/10 修正；教育部 2024/8/26 修正)。漢字承載著客語的歷史文化內涵，並將其與更廣泛的漢語語言傳統聯繫起來。

大埔(Dapu) 語者人數						
年齡	~17	18~30	31~50	51~65	66~	合計
男性	9	5	12	20	20	66
女性	7	9	21	32	10	79
合計	16	14	33	52	30	145
詔安(Zhao'an) 語者人數						
年齡	~17	18~30	31~50	51~65	66~	合計
男性	2	3	7	23	16	51
女性	7	9	9	18	12	55
合計	9	12	16	41	28	106

表 2. HAT 語料庫中大埔腔和詔安腔 ASR 子集錄音語者的年齡和性別分佈。

除了正規文字以外，教育部也制定了客語音標，採用拉丁字母，稱為「臺灣客語拼音方案」(教育部 2024/8)。此拼音方案由教育部於 2012 年推出，旨在規範客家語發音的拼音，使不熟悉漢字客語發音的人更容易使用。

表 1 展示了六種腔調之臺灣客語拼音聲調。臺灣客語拼音方案中的聲調符號有兩種：調型與調值。調型類似注音符號中的聲調符號，有「ˊ」、「ˇ」、「ˋ」、「^」、「+」等；調值則是以阿拉伯數字，如 24、53，表示聲調的高低變化。聲調無論調型或調值，統一採音節右上標。例：客家 hagˋ gaˊ/ hag^{2} ga^{24}。上表中的 v 表舒聲韻音節；vd 表示入聲韻音節。大埔、詔安腔入聲調下加底線，表調值之短促，如：21、54。

教育部臺灣客語辭典的聲調標記，除了調型與調值外還有調號。客語調號和台語調號一樣，都是對應傳統聲調調名－平、上、去、入的序號，且分成陰陽兩組：調號 1~4 對應陰平~陰入；調號 5~8 代表陽平~陽入。不同腔調的客語，部分聲調種類會合併，因此實際上的聲調種類只有 6 或 7 種。其中，大埔腔另有超陰平調（調值 33），調號為 9，是陰平調的一種特殊變調。另外，南四縣腔部分地區（如美濃）之陰平調調值 33，與標準調值 24 不同，調型以「+」表示，但調號仍是 1。

在 HAT 語料庫與 FSR 競賽資料集中，聲調拼音一律採用調值表示，且不遵循客語拼音方案的上標表示規定，以方便電腦處理。但應當注意的是，不同腔調的相同調值，不一定代表相同的聲調種類。例如：調值 53，在海陸腔是陰平、在詔安腔是陽平、在饒平

音檔	大埔	詔安	合計
音檔筆數	75,239	98,119	173,358
語音句數	151,719	179,164	330,883
時分秒數	204:10:55	213:34:03	417:44:58

表 3. HAT 語料庫中大埔腔和詔安腔 ASR 子集語音資料的數量和長度統計。

音檔	大埔男	大埔女	詔安男	詔安女	合計
音檔數	17,592	16,012	17,616	16,661	67,881
語句數	37,701	34,473	35,773	33,852	141,799
時分數	31:30	30:21	32:25	31:27	125:44

表 4. HAT 語料庫中大埔腔和詔安腔 TTS 子集語音資料的數量和長度統計。

腔是上聲、在大埔腔則是陰去聲。所以，這會使得混合多種腔調的語料進行語音辨識模型訓練時，在拼音辨識上會變得更複雜。

3 HAT 臺灣客語語音資料庫

HAT 語料庫包含兩類錄音語料子集：自動語音辨識 (ASR)、文字轉語音合成 (TTS)。關於 ASR 與 TTS 子集的錄音設計、錄製方式，以及四縣腔與海陸腔的語料統計請參考 (Liao, 2023)。基本上大埔腔與詔安腔語料的錄製方式與四縣與海陸一致，主要差別在於合適的錄音人比較難找，所以蒐集到的語料之語者人數較少。另外語者的年紀也偏大，因此將年齡統計 50 歲以上再細分為 65 歲上下兩群。

在 ASR 子集的部分，經過兩年的努力，我們招募了 251 名錄音語者，共錄製了 330,883 句語音，相當於 417 小時的客家話語音，涵蓋了大埔和詔安兩個腔調。表 2 詳細列出了最終完成的大埔和詔安語音辨識 ASR 子集中說話者的年齡和性別分佈。從中可看出超過三分之一的語者年齡在 51~65 歲。表 3 則分別列出了大埔和詔安 ASR 子集中的音檔筆數、語音句數、和音檔時間總長的時分秒數。每個音檔可能包含多句語音，而其分句是根據錄音提示稿的標點符號。

在 TTS 子集的部分，每個腔調由一名男性和一名女性母語者錄製。基於 TTS 模型訓練的需求，每位錄音語者各自錄製了約 30 小時的語音。表 4 顯示了四名 TTS 錄音語者的音檔統計數據，包含音檔的檔案數、語句數、和長度時分數。

上述 HAT 語料庫中的每個語音檔都帶有相應的 JSON 格式元數據 (metadata)，其中包括錄音提示卡編號、音檔編號、音檔長度、提示句客語漢字與拼音及其華語翻譯，以及說話人特徵和錄音環境等資訊。例如：

```
DF1010001J2003_1.json
{
  "提示卡編號": "J2003"
  "音檔編號": 1
  "發音員編號": "DF101"
  "音檔長度": "00:04"
  "客語漢字": "五月節愛食粽。"
  "客語拼音": "ng31 ngied54 zied21 oi33
               shid54 zung53 。"
  "華語字": "端午節要吃粽子。"
  "性別": "女性"
  "年齡": 72
  "身分別": "薪傳師（教師）"
  "現居地": "臺中市東勢區"
  "18 歲前居住地": "臺中市東勢區"
  "教育程度": "大學"
  "錄音腔調": "大埔"
  "錄音環境": "一般辦公室"
  "流暢度": "普通"
}
```

4 FSR-2025 - ASR II 客語語音辨識競賽

為了推廣 HAT 語料並評估先進客語自動語音辨識系統的性能，FSR-2025 挑戰賽於 2025 年 6 月 2 日至 10 月 6 日舉行，歡迎學術界和產業界組隊參加。最後共有 20 隊報名，並有 16 隊完賽，包含學生組 13 隊、社會組 3 隊。

FSR-2025 挑戰賽的主要任務是建造一個能辨識臺灣大埔腔與詔安腔客語之語音辨識器，該任務根據語音辨識輸出分為兩個賽道 (Track)：

(1) Track1: 臺灣客語漢字—語音辨識輸出漢字
例如：今晡日係拜二。
辨識率計算漢字字元錯誤率 (CER)。

(2) Track2: 臺灣客語拼音—語音辨識輸出拼音
例如：gim24 bu24 ngid2 he55 bai55 ngi55.
辨識率計算拼音音節錯誤率 (SER)。

挑戰賽共進行兩次評測：熱身賽、決賽。熱身賽的目的是讓參賽者先進行一次模擬測試，以驗證其系統，故成績僅供參考。最後的決賽成績才做為競賽的結果與排名。

	語者數	語句數	字元數	時數	
total	**> 51**	**36,316**	**542,617**	**82.23**	
讀稿	17	12,197	180,055	31.43	大
讀稿	16	15,152	199,870	30.59	詔
Train	**33**	**27,349**	**379,925**	**62.02**	
讀稿	3	1,304	18,639	4.00	大埔腔
即席		445	11,877	1.08	
小計	> 3	1,749	30,516	5.08	
讀稿	6	2,154	26,105	4.00	詔安腔
即席		501	15,006	1.13	
小計	> 6	2,655	41,111	5.13	
Eval	**> 9**	**4,404**	**71,627**	**10.21**	
讀稿	5	1,122	15,390	2.97	大埔腔
即席		904	27,496	2.03	
小計	> 5	2,026	42,886	5.00	
讀稿	4	1,322	15,119	2.57	詔安腔
即席		1,215	33,060	2.43	
小計	> 4	2,537	48,179	5.00	
Test	**> 9**	**4,563**	**91,065**	**10.00**	

表 5. FSR-2025-Hakka II 語音資料集。

配合競賽的時程，依序釋出下列競賽用語音資料集（相關統計參見表 5）：

- 訓練集 (FSR-2025-Hakka-**Train**)：來自 HAT 語料庫約 60 小時、33 位語者的語料，包含大埔腔和詔安腔約各半的語料和語者。參賽者報名後即可獲得此訓練集。
- 驗證集 (FSR-2025-Hakka-**Eval**)：這是在熱身賽階段釋出作為驗證測試之用。如表 5 中所示，除了來自 HAT 語料庫的讀稿 (read) 錄製 (Record) 語料外；為了提高語音辨識難度，我們另外增加了從媒體 (Media) 節目中擷取的即席 (spontaneous) 語音。驗證集語音總時數約為 10 小時，兩腔調各 5 小時，讀稿和即席語音的時數大約為 4 比 1。
- 測試集 (FSR-2025-Hakka-**Test**)：這是用在決賽中的測試語音。與熱身賽驗證集大小一樣都是 10 小時，且也都有來自 HAT 錄製的讀稿語音和來自媒體的即席語音，但是兩者的比例約為 1 比 1。因為較難辨識的即席語音比例提高，因此決賽的挑戰更大。

隊伍編號	漢字 CER	拼音 SER	漢字名次	拼音名次	幾何平均
U	6.84%	19.57%	1	3	2
M	8.84%	13.44%	3	1	2
D	10.51%	14.72%	8	2	4
P	7.64%	89.49%	2	13	5
T	10.01%	47.14%	4	10	6
H	15.92%	20.49%	10	4	6
B	10.42%	23.40%	7	6	6
F	17.92%	21.91%	12	5	8
R	10.33%		5	*17*	9
E	16.06%	29.09%	11	8	9
J	10.37%		6	*17*	10
K	13.36%	70.68%	9	12	10
O	101.63%	25.52%	17	7	11
C	37.64%	35.75%	15	9	12
L	31.93%	51.47%	13	11	12
G	75.58%	100.97%	16	14	15
N	36.68%		14	*17*	15

表 6. 熱身賽辨識結果與排名。

	讀稿	即席	綜合	大埔	詔安
漢字 CER	1.82%	15.03%	**6.84%**	3.96%	8.99%
拼音 SER	6.51%	21.66%	**13.44%**	9.02%	16.75%

表 7. 熱身賽各種語料類別之最佳辨識結果。

4.1 熱身賽結果

表 6 列出熱身賽參賽 16 隊與主辦單位建置之基準 (Baseline) 系統（隊伍編號 B），共 17 組針對驗證集 (FSR-2025-Hakka-**Eval**) 之辨識結果與排名。總排序是依據漢字排名和拼音排名的幾何平均，若平均相同則依據漢字排名。隊伍編號除 B 以外，A~P 為學生組，Q~U 為社會組。最後 A, I, Q, S 四隊未提交結果參與評測，故不予排名。剩下參賽 16 隊中有 3 隊未提出拼音辨識結果，故假定其拼音排名並列 17。Track1 漢字賽道參賽 16 隊中有 6 隊成績優於基準；Track2 拼音賽道參賽 13 隊中則有 5 隊成績優於基準。總排序則共有 6 隊在基準系統 B 前面。

表 7 顯示最佳結果是 Track1 漢字 CER = 6.84%；Track2 拼音 SER = 13.44%。顯然漢字的辨識率優於拼音，錯誤率相差一倍。若依據語音性質分類，可以看到即席語音辨識率遠低於讀稿語音，這也是因為訓練語料中只有讀稿語音。另外若依據腔調分類，則可以

隊伍編號	漢字 CER	無調拼音 SER	拼音 SER	漢字名次	拼音名次	幾何平均
E	7.50%	17.45%	25.04%	1	6	2
P	8.99%	12.36%	19.22%	2	3	2
D	11.21%	11.32%	15.08%	4	2	3
M	22.50%	10.49%	14.81%	13	1	4
U	9.46%	20.97%	30.44%	3	8	5
H	15.73%	13.82%	20.68%	7	4	5
B	17.13%	14.43%	23.50%	9	5	7
L	16.05%	18.97%	27.57%	8	7	7
O	15.61%	26.60%	35.20%	6	11	8
J	13.35%			5	*17*	*9*
S	18.78%	21.30%	33.38%	10	10	10
N	28.70%	20.70%	30.45%	15	9	12
K	21.21%	37.80%	47.95%	12	12	12
R	19.08%			11	*17*	*14*
C	30.40%	56.67%	59.35%	16	13	14
F	26.43%			14	*17*	*15*
G	503.0%	503.4%	503.4%	17	14	15

表 8. 決賽辨識結果與排名。

看到大埔腔的結果優於詔安腔，其中原因尚待探究。

4.2 決賽結果

表 8 列出決賽參賽 16 隊與基準系統 B，共 17 組針對測試集 (FSR-2025-Hakka-**Test**) 之辨識結果與排名（注意：T 隊只參加熱身賽；S 隊只參加決賽）。有 3 隊沒有提出拼音辨識結果，故假設其拼音排名並列 17。共有 6 隊平均排名領先基準系統 B，其中編號 E 的隊伍成績進步最大，從熱身賽平均第 9 名進步到決賽時並列最優的第 2 名；其它 5 名在熱身賽時也排在基準系統 B 之前。當中兩次拼音排名第一的 M 隊，其漢字成績異常，名次從第 3 掉到 13，可能是決賽的系統有問題。

決賽額外計算了「無調拼音 SER」，就是拼音結果不管聲調對錯，只計算非聲調部分的音節錯誤率。從結果來看，拼音不管聲調後 SER 明顯進步很多，除了漢字成績異常的 M 以外，還有 H, B, N 這三組的無調拼音 SER 低於漢字 CER。這可能表示目前混合大埔與詔安腔的辨識模型，在聲調方面採取調值表示會產生混淆問題；而且目前都是標示本調，

編號	CER SER	名次	賽道	主要特色	Whisper	variant model	語料 1	語料 2	TTS	DA	DA Robust strategy
E	**7.5%**	1	漢	兩階段 FT 全參數 FT	Large-v2		HAT-Vol-1	FSR-train		SpecAugment	
	25.0%	6	拼	後處理：①二腔漢字 G2P 比較拼音，取 WER 小者②字典比對修正	同上		同上	FSR-train → 漢字 Encoder + 拼音 Decoder		同上	
P	9.0%	2	漢	dialect-aware special tokens	Large-v3	FormoSpeech Hakka	FSR-train FSR-eval		VoxHakka VC speaker DA	SpecAugment -freq. masking -time masking	Progress. Augm. : Speed+Spec +Noise
	19.2%	3	拼	同上	同上		同上		同上	同上	同上
D	11.2%	4	漢	降噪前處理	Medium		FSR-train FSR-eval	LoRA 效果不佳		SpecAugment Audio Concat.	
	15.1%	2	拼	同上	同上		同上	同上		同上	
M	22.5%	13	漢	N-best 文本候選 mBART 文本修正 x5 RNNLM 重評分	Large-v3 +LoRA	3 種模型 (大埔/詔安/混合)輸出 5-best 候選句	FSR-train FSR-eval 處理合音字+刪錯讀	DA x6 倍		①MUSAN Noise SNR ∈{5,10,15}dB ②Speed perturb.	
	14.8%	1	拼	腔調辨識+腔調 ASR 模型 RNNLM 重評分	Medium	2 種模型 (大埔/詔安) 2 選 1	同上	同上		同上	
U	9.5%	3	漢		Large-v3 +AdaLoRA	FormoSpeech Hakka	FSR-train FSR-eval	詔安 OOV (TTS)	formospeech /yourtts-htia-240704		速度、音高變動與空氣吸收
	30.4%	8	拼	Kaldi K2 模型	n/a	Wenet-Zipformer	HAT-Vol-1 = 四 396h + 海 300h	FSR-2023 FSR-train FSR-eval	同上		
H	15.7%	7	漢		Large-v3		FSR-train FSR-eval	Radio 11hr E-learning 16hr	VITS: 辭典例句+Media 語料	Static: MUSAN, MetricAug	Dynamic: Audiomentations
	20.7%	4	拼		同上		同上			同上	同上
B	17.1%	9	漢		Large-v3 Turbo		FSR-train	串接音檔 159.5 小時		Audio Concat. SpecAugment	Audiomentations
	23.5%	5	拼		同上		同上	同上		同上	同上

表 9. 前六名領先隊伍與基準系統 (編號 B) 的主要作法摘要。

但兩種腔調有各自不同的變調規則，混為一體恐難以訓練建模。

最後，比較表 8 與表 6 的結果，可以看出相對於熱身賽階段，決賽時領先群隊伍的系統都有進步。因為雖然決賽測試集的難度比熱身賽驗證集的難度要高不少（即席語音比例從 20%增加到 50%），但最佳成績差異不大。決賽最佳結果分別是 Track1 漢字 CER = 7.50%；Track2 拼音 SER = 14.81%。這當中的改進至少應該包含模型訓練增加了熱身賽驗證集 (FSR-2025-Hakka-**Eval**)的語料，尤其當中所含的媒體語料是原來訓練集中完全欠缺的即席語音類型。

4.3 領先隊伍作法摘要

表 9 是從參賽隊伍投稿的論文中，簡單摘要出領先六隊跟基準系統的主要作法。除了 U 隊在拼音賽道採用 Kaldi K2 模型外，全部參賽系統都是基於 Whisper 核心模型的各種版本。P、U 兩隊的漢字賽道使用 FormoSpeech 公開的模型 whisper-large-v3-taiwanese-hakka (FormoSpeech, 2025) 做為基底，此模型是將 Whisper Large-v3 經過六種客語腔調的微調訓練而得。其餘大部分系統都直接使用 Whisper Large-v3，只有兩隊使用 Medium、一隊使用 Large-v2，唯有基準系統採用 Large-v3 Turbo。

若觀察漢字賽道第一名的 E 隊，其主要特色是以兩階段微調訓練 Large-v2 模型：第一階段訓練語料是 2023 年釋出的 HAT-Vol-1 語料庫，包含四縣腔 396 小時、海陸腔 300 小時；第二階段訓練語料是大會提供的大埔腔與詔安腔語料—FSR-train 62 小時、FSR-eval 10 小時。另外也簡單提到使用 SpecAugment (Park et al., 2019) 進行資料增強。他們的模型則是使用 ESPNet toolkit (Watanabe et al., 2018) 進行實現。

漢字第二名 P 隊與第三名 U 隊辨識率接近，兩隊的作法也有很多相似之處，都是使用 FormoSpeech 的客語基礎模型，也都使用 TTS 來擴充訓練語音。但 P 隊的特色是針對 Whisper 設計特殊 token 來辨識腔調，包括大埔腔、詔安腔、與未知腔，藉此讓模型對腔調有意識的進行區分來影響語音辨識結果。另外與第一名相同，P 隊也採用 SpecAugment 進行資料擴增，U 隊則無。

在拼音賽道方面，第一名 M 隊與第二名 D 隊的辨識率相近；兩隊的共通點就是都採用較小的 Whisper Medium 模型。M 隊的特色就是額外訓練一個 RNNLM 語言模型（漢字採用 LSTM、拼音採用 GRU），對於 Medium 模型輸出的 10-best 候選句來進行重評分；另外就是利用漢字辨識結果進行斷詞與查辭典，藉以辨識腔調種類，再以腔調專屬拼音辨識模型來進行辨識。此外，M 隊對於大會提供的語料也進行過濾，刪除備註語者錯讀的音檔，也處理合音字的特殊表示。

拼音第二名 D 隊的特色則是對語音進行降噪前處理，包括訓練語音和測試語音；另外也採用 SpecAugment 和串接短音段來進行訓練語料擴充。拼音第三名是 P 隊，其辨識率跟前兩名差距較大。P 隊的拼音作法幾乎與漢字作法完全一樣，只差在拼音是使用 Large-v3 原生模型，而不是 FormoSpeech 的客語模型。

5 結論與展望

本文不僅回顧了 HAT 語料庫第二階段中的大埔腔與詔安腔子集，也介紹了針對這兩種腔調的 FSR-2025 客語語音辨識挑戰賽。HAT 計畫也將持續錄製剩餘的客家方言。由於目前的 HAT 語料庫專注於讀稿語音，因此也開始著手轉錄來自電視、廣播等傳統媒體以及 YouTube、播客等線上平台的自發性即席語音，尤其是訪談形式的語音。最終目標是收集足夠的台灣自發性客家話語音數據，藉以建立更先強韌可靠的客家語自動語音辨識系統。

Acknowledgments

This work was supported in part by the Co-creation Platform of the *Speech-AI Research Center (SARC)*, *Industry-Academia Innovation School (IAIS)*, NYCU, under the framework of the National Key Fields Industry-University Cooperation and Skilled Personnel Training Act, from the Ministry of Education (MOE), the National Development Fund (NDF), industry partners in Taiwan, and in part by the *Hakka Affairs Council* (客委會).

References

FormoSpeech. 2025. whisper-large-v3-taiwanese-hakka. https://huggingface.co/formospeech/whisper-large-v3-taiwanese-hakka. Accessed: 2025-09-10.

Y. -F. Liao et al. 2023. Taiwanese Hakka Across Taiwan Corpus and Formosa Speech Recognition Challenge 2023 - Hakka ASR. *26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Delhi, India, 2023, pp. 1-6, doi: 10.1109/O-COCOSDA60357.2023.10482979.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, pages 2613–2617.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. In *Interspeech 2018*, pages 2207–2211.

主計總處 2020：表６、６歲以上本國籍常住人口使用語言情形。*109 年人口及住宅普查>綜合報告統計表*。Accessed: Nov. 2, 2025. [Online]. https://www.stat.gov.tw/News_Content.aspx?Create=1&n=2755&state=1327FD6AD8DCDA52&s=230300&ccms_cs=1&sms=11065

客委會 2024：臺灣客語語音資料庫(HAT)。*中華民國計算語言學學會/語料庫*。https://www.aclclp.org.tw/use_mat_c.php#hat。

教育部 2024/8：臺灣客語拼音方案使用手冊。*教育部語文成果網/字音類/臺灣客語拼音*。

Accessed: Nov. 2, 2025. [Online]. Available: https://language.moe.gov.tw/index.aspx

教育部 2024/8/26 修正：臺灣客語書寫推薦用字（第 2 批）。*教育部語文成果網/字形類/臺灣客語用字*。Accessed: Nov. 2, 2025. [Online]. Available: https://language.moe.gov.tw/index.aspx

教育部 2025/7/10 修正：臺灣客語書寫推薦用字（第 1 批）。*教育部語文成果網/字形類/臺灣客語用字*。Accessed: Nov. 2, 2025. [Online]. Available: https://language.moe.gov.tw/index.aspx

陽明交大 2023："Formosa Speech Recognition Challenge 2023 - Hakka ASR." Accessed: Aug. 23, 2023. [Online]. Available: https://sites.google.com/nycu.edu.tw/fsw/home/challenge-2023

陽明交大 2025："Formosa Speech Recognition Challenge 2025 - Hakka ASR II." Accessed: Aug. 23, 2023. [Online]. Available: https://sites.google.com/nycu.edu.tw/fsw/home/challenge-2025

賴維凱 2008：臺灣客家話的分布及使用概況。*教育部臺灣客語辭典/客語資源/客語知識庫/綜論*。Accessed: Nov. 2, 2025. [Online]. Available: https://hakkadict.moe.edu.tw/resource/

低資源語言的語音辨識：客語漢字與拼音模型比較
Speech Recognition for Low-resource Languages: A Comparative Study on Hakka Han Characters and Romanization

鄭宇翔
Department of Computer Science and Information Engineering, National Taitung University / Rm. SEC408, No. 369, Sec. 2, University Road, Taitung City, Taitung County 950309, Taiwan
karlboy3306@gmail.com

吳亦軒
Department of Computer Science and Information Engineering, National Taitung University / Rm. SEC408, No. 369, Sec. 2, University Road, Taitung City, Taitung County 950309, Taiwan
pu1254@gmail.com

摘要

本研究針對低資源語言的語音辨識，以客語為例進行探討。由於目前缺乏專門處理閩南語、客語及原住民族語的語音模型，本研究以 OpenAI Whisper-Medium 為基礎，並透過 LoRA（Low-Rank Adaptation）進行微調，建立兩種不同輸出形式的模型：客語漢字與客語拼音模型。實驗資料共計約 80 小時，涵蓋大埔腔與詔安腔，並分別以字元錯誤率（CER）與詞錯誤率（WER）評估模型表現。

Abstract

This study focuses on speech recognition for low-resource languages, with Hakka as the case study. Since there is currently a lack of dedicated speech models for Taiwanese Southern Min, Hakka, and indigenous languages, we adopt OpenAI Whisper-Medium as the base model and apply Low-Rank Adaptation (LoRA) for fine-tuning. Two models with different output forms were developed: a Hakka character-based model and a Hakka phonetic-based model. The experimental dataset contains approximately 80 hours of speech, covering the Dapu and Zhao'an dialects, and the models were evaluated using Character Error Rate (CER) and Word Error Rate (WER).

關鍵字：低資源語言、客語、語音辨識
Keywords: Low-resource Languages, Hakka, Speech Recognition

1 緒論

在語音辨識的研究中，高資源語言（如中文與英文）已達到相當高的準確度，而這些語言擁有大規模的語音與文字的對照語料庫，以及成熟的自然語言處理資源，相較之下，低資源語言由於缺乏大規模語料及相關工具，研究與應用的進展相對有限，導致語音辨識的效果普遍不佳。

近年來，隨著 OpenAI Whisper 等大型多語言語音模型的出現，研究者開始嘗試利用遷移學習與微調（fine-tuning）技術，將這些模型應用於低資源語言，以彌補語料不足的缺陷。Whisper 此種多語言模型使其能在缺乏資料的情況下，仍展現出一定程度的泛化能力，為低資源語言的語音辨識研究提供了新的可能性。然而，如何設計合適的標註策略與輸出格式，仍是提升辨識效能的重要議題。

對於聲調語言而言，標註方式的選擇非常關鍵，以客語為例，其並沒有統一的書寫標準：一方面可以使用漢字進行書寫，另一方面也能以羅馬拼音搭配數字標註聲調的方式呈現。兩種標註系統各具優缺點：漢字符合使用者的閱讀習慣，但存在多音字與語音與文字對應不一致的挑戰；拼音則能直接反映語音特徵，減少歧義，卻可能因使用者不熟悉而降低應用價值。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 435–440
November 20-22, 2025 ©2025 Association for Computational Linguistics

本研究以客語為例，探討在相同語音辨識架構下，分別使用客語漢字與客語拼音作為輸出標註，對模型效能所造成的差異。我們以 OpenAI Whisper-Medium 為基礎，並透過 LoRA（Low-Rank Adaptation）進行微調，建立兩種模型，並分別以字元錯誤率（Character Error Rate, CER）與詞錯誤率（Word Error Rate, WER）進行評估。透過比較兩種標註策略的實驗結果，我們期望提出一套適用於低資源語言的有效訓練流程，並提供對未來客語與其他低資源語言語音辨識研究的參考。

2 相關研究

A. 低資源語言 ：低資源語言的語音辨識研究受到廣泛關注。(江宥呈,2023)提出 VoxCentum 資料集涵蓋了 137 種語言共 13,072 小時語音，指出資料集不平衡會顯著影響模型效能，而平衡語料與對比學習能有效提升泛化能力。(劉廷緯 ,2024) **提出了低資源語言的語音處理**，特別是如何在資料不足的情況下，利用 **自監督式學習 (self-supervised learning, SSL)** 來提升語音辨識 (ASR) 與語音處理效能。這些研究顯示低資源語言不僅依賴語料量，也需要設計合適的訓練與增強策略。本研究延續此方向，進一步探討 標註格式 (漢字 vs. 拼音) 對模型效能的影響。

B. Whisper 模型：OpenAI 所提出的 Whisper 模型，已成為多語言語音辨識的重要基礎。(呂可名,2024)基於 Whisper 開發即時語音辨識與語者分段系統，驗證了其在多人對話與多語境的強大適應性。(Hsieh et al.,2023) 則針對台語與中文進行 Whisper 微調，利用 Common Voice 與台語戲劇資料，共約 800 小時語料，最終 CER 約 50.7%，顯示 Whisper 在低資源語言上的潛力，但仍需更多資料與後處理。另一項研究比較 Whisper 與 Wav2vec2 在台語辨識的表現，發現 Whisper 在跨語言適應上具優勢，但仍面臨書寫系統不一致的挑戰。這些研究突顯了 Whisper 在 多語言與低資源語言 環境下的強大泛化能力，也為本研究比較「漢字 vs. 拼音」提供了方法論上的基礎。

3 方法

3.1 語料

本研究使用 FSW Challenge 2025 所公開的客語語料，涵蓋大埔腔與詔安腔兩種主要方言，總長度約 80 小時。每筆語料均附有兩種標註，此設計為我們提供了直接比較不同標註策略的可能性，並可探討文字表示對語音辨識效能的影響。語料經過整理後，依照 8:1:1 的比例劃分為訓練集（80%）、驗證集（10%）、測試集（10%），並確保兩種腔調的比例在各資料集內保持平衡，以避免模型因資料分布不均而產生偏差。值得注意的是，雖然主辦方分別提供了約 40 小時的大埔腔與 40 小時的詔安腔語料，但本研究並未將兩者分開訓練，而是統一整合後進行模型訓練。此設計的原因在於：若模型僅依賴單一腔調語料可能會導致模型對特定腔調過度擬合，進而降低對其他腔調的辨識效果，透過將不同腔調混合訓練，模型能同時學習多樣化的發音特徵提升其泛化能力，使其在實際應用中面對不同腔調輸入時，仍能維持穩定的辨識表現。

3.2 音訊前處理

為確保資料一致性，所有音檔在訓練前均進行以下處理：

- 單聲道轉換：將立體聲檔案轉為單聲道，以降低計算負擔。
- 重取樣：將音訊取樣率統一至 16 kHz，與 Whisper 模型的輸入規範一致。
- 峰值正規化：對所有音檔進行正規化，以避免因音量差異過大導致訓練不穩定。

此外，本研究在訓練過程中加入 輕量級資料增強技術(data augmentation)，以模擬多樣化的語音環境，提升模型泛化能力:

A. 音高偏移(Pitch Shifting)：在不影響語義的情況下，對音訊進行小幅度隨機音高調整，使模型能夠學習到不同人說話及語境下的聲學變化，此方法能增加語音的多樣性，尤其對於有限語料的低資源語言來說，有助於提升模型的泛化能力。

B. 雜訊注入(Noise Injection)：在音訊中加入低訊噪比的高斯雜訊，以模擬真實場景中可能出現的背景噪音。由於實際應用環境（如會議、課堂、日常對話）常存在干擾聲，本研究透過此方法使模型能夠學習在雜訊下仍保持穩定辨識能力。

C. 音量縮放(Volume Scaling)：機將音訊振幅調整至原本的 0.8 至 1.25 倍，模擬不同錄音設備、錄音距離或說話音量的差異，避免模型對固定音量過度擬合，進而提升其對不同輸入條件的魯棒性。

與部分研究常見的語速變化(Speed Perturbation)不同，本研究刻意避免使用此方法。原因在於 Whisper 採用固定時間解析度的聲譜表示，若對語料進行語速改變，可能導致資料分布偏離模型的原始特徵空間，進而影響訓練穩定性，因此本研究以音高偏移作為主要的增強手段。

3.3 模型架構與 LoRA 微調

本研究以 OpenAI Whisper-Medium 模型作為基底。Whisper 是一種基於 Transformer 編碼–解碼器架構的多語言語音辨識模型，具備跨語言的強大泛化能力，特別適合用於低資源語言的研究。然而直接微調完整模型需要大量運算資源，因此本研究採用 Low-Rank Adaptation (LoRA) 技術進行參數高效化的調整。LoRA 的優點在於僅需訓練少量附加參數，顯著降低訓練成本，同時保留模型對其他語言的泛化能力。

在此基礎上，本研究設計了兩組實驗模型：

1. 漢字模型：輸出客語漢字，學習率設定為 5e-5，訓練 10 個 epoch。選擇較低學習率與較長訓練週期，目的是讓模型在有限語料下能更穩定地擬合字元級輸出。
2. 拼音模型：輸出帶數字聲調的拼音，學習率設定為 1e-3，訓練 5 個 epoch。由於拼音單位較漢字單純，模型較容易收斂，因此選擇較高的學習率與較短的訓練週期，以加速收斂並避免過擬合。

3.4 評估指標

為比較不同標註策略，本研究設計兩種實驗：Track 1（漢字）：以字元錯誤率 (CER) 作為評估指標，並計算模型輸出與標註的差異。Track 2（拼音）：以音節錯誤率 (SER) 作為評估指標，計算模型輸出與標註在音節的差異。計算公式如下：

$$錯誤率 = \frac{S + D + I}{N}$$

其中：

S：替換數（模型輸出錯誤的單位數）

D：刪除數（模型輸出缺少的單位數）

I：插入數（模型輸出多餘的單位數）

N：參考標註的總單位數

在 Track 1（CER）中，單位為 漢字字元；在 Track 2（SER）中，單位為 帶有數字調的拼音音節。

範例（CER）

參考標註（漢字）：「客語」（共 2 個字）

系統輸出：「語」（1 個字）

$$S = 0, D = 1, I = 0, N = 2$$

$$CER = (0 + 1 + 0)/2 = 0.5\ (50\%)$$

範例（SER）

參考標註（拼音）：ng31 ngied54（2 個音節）

系統輸出： ng31 ngid54（2 個音節，第二音節調號錯誤）

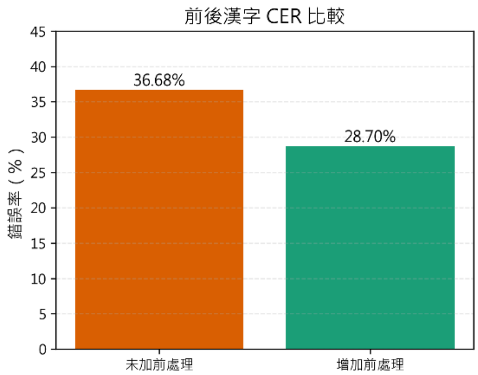
$$S = 1, D = 0, I = 0, N = 2$$

$$SER = (1 + 0 + 0)/2 = 0.5\ (50\%)$$

4 實驗結果

4.1 客語漢字模型結果

本研究首先針對客語漢字模型在未進行前處理時進行測試，計算整體字元錯誤率(Character Error Rate, CER)。結果顯示，模型的 CER 為 36.68%。在經過前處理後 CER 下降到了 28.70% 。(圖一)



(圖一)

CER 仍偏高的原因我們發現為以下兩點:

A. 多音字現象

客語中存在大量的多音字現象，即同一漢字對應多個不同的發音與語義。例如：

- 「著」可讀作 *tok5*（表示「穿著」）或 *zok8*（表示「正在」）;
- 「會」可讀作 *voi5*（能夠）或 *hoi5*（開會/聚會）。在語音辨識中，模型需從語音特徵正確選擇對應的漢字，但由於上下文有限以及語料不足，模型常會出現替換錯誤。例如，輸出「正在」時，可能誤判成「穿著」，造成 CER 提升。相較之下，拼音標註方式能更精準地對應聲學特徵，避免了多音字歧義。

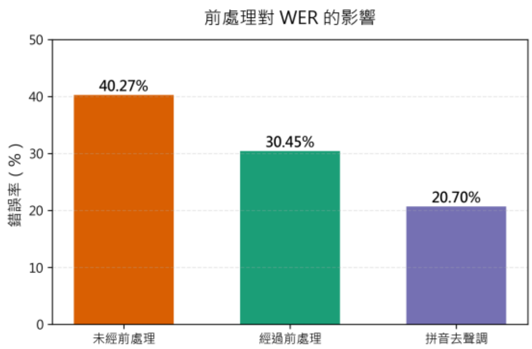
B. 語料規模有限

本研究所使用的語料訓練總量約為 60 小時，對於現代深度學習語音模型而言仍屬於小規模資料。雖然語料已涵蓋大埔腔與詔安腔，但在詞彙分布上仍顯不足：

- 常見詞：如「食」、「飲」、「行」等詞彙模型能學習得較好；
- 稀有詞彙：例如專有名詞、方言特殊用語，出現頻率極低，導致模型在遇到測試集中的新字時錯誤率升高。這種情況尤其會影響漢字模型，因為字表龐大（漢字數量遠多於拼音單位數量)，有限的資料無法充分涵蓋所有字形，導致模型對少見字的預測準確率明顯下降。

4.2 客語拼音模型結果

本研究接著針對**客語拼音模型**進行測試，計算整體**詞錯誤率** (Word Error Rate, WER)。拼音模型未經前處理的 **WER 為 40.27%**，而經過前處理後的 WER 為 30.45，拼音去聲調的 WER 為 20.70。(圖二)



(圖二)

以下整理出我們觀察到的錢處理過後顯著的影響：

A. 資料前處理的重要性

前處理包含語音正規化、去除異常符號及資料一致化，能夠有效減輕模型受到資料雜訊的影響，讓訓練更集中於語音與拼音對應關係。拼音去聲調後 WER 顯著下降，顯示聲調雖具語義區分作用，但若模型尚未能正確捕捉其音韻特徵，反而會降低辨識準確度。這意味後續研究可針對聲調特徵進行專門建模，如增加 tone embedding 或 tone-aware acoustic feature。

B. 字詞規模較小

漢字的字表可能高達數千甚至上萬個字，對低資源語料而言，許多字在訓練集中出現頻率極低，模型難以學習。相對而言，拼音的音節組合有限。例如，客語的聲母、韻母與聲調的組合數量遠少於漢字總量，詞彙表規模縮小至數百個單位即可涵蓋主要發音。這樣的差異讓拼音模型在訓練時更容易收斂，並在測試階段遇到陌生語音時仍能正確對應到既有音節單位，降低了替換錯誤與刪除錯誤的機率。

4.3 漢字與拼音模型比較

綜合兩種標註策略的結果我們可得知以下觀察：

1. **效能比較**：拼音模型 (WER 20.70%) 明顯優於漢字模型 (CER 28.70%)。這顯示拼音作為中介表示更貼近聲學特徵，有助於模型學習與收斂。
2. **實用性比較**：雖然拼音模型效能更佳，但輸出結果對一般使用者不直觀，閱讀成本高；相反地，漢字模型雖然錯誤率較高，但輸出內容更符合使用習慣，應用潛力較大。

5 結論與未來展望

5.1 結論

本研究以客語為例，探討低資源語言語音辨識中不同標註策略對模型效能的影響。我們基於 OpenAI Whisper-Medium 模型，透過 LoRA 微調建立兩種模型：輸出客語漢字與客語拼音的模型。實驗結果顯示：

1. **漢字模型** 的字元錯誤率 (CER) 為 **28.70%**，顯示在文字輸出上仍受限於書寫不一致、多音字現象以及語料不足。
2. **拼音模型** 的詞錯誤率 (WER) 僅為 **20.70%**，效能顯著優於漢字模型，因為拼音標註與聲學特徵的直接對應、詞彙表規模小等原因。
3. 雖然拼音模型在效能上優勢明顯，但漢字模型在應用層面更具可讀性與實用性，因此這兩種策略各有優缺，未來也應考慮整合多模型以同時兼顧準確率與使用者需求。

5.2 未來展望

基於上述研究成果，我們提出以下未來方向：

1. **拼音轉漢字模組**：結合拼音模型與漢字轉換系統，透過語言模型或字典資源進行後處理，提升輸出的可讀性。
2. **跨腔調擴展**：納入更多客語方言（如四縣腔、海陸腔），驗證模型在多樣化腔調下的泛化能力。
3. **語料擴增**：蒐集更大規模的客語語音與標註，並透過自動增強技術 (如:非監督學習) 補足現有不足。
4. **模型比較與優化**：嘗試更小或更大的 Whisper 模型版本，以及其他低資源語言專用架構，進一步驗證效能差異。
5. **應用場景實驗**：將模型部署於真實應用，如客語教學平台、語音輸入法或語言保存工具，檢驗其實際效益與使用者接受度。

6 參考文獻

A. Liu, W. (2025). *Enhancing Efficiency and Reliability in Automatic Speech Recognition Systems* (Doctoral dissertation, The Chinese University of Hong Kong (Hong Kong)).

B. 陳昇德. (2025). 基於大型語音模型的模型壓縮技術應用於邊緣運算裝置. *淡江大學機械與機電工程學系碩士班學位論文*, 1-48.

C. 龙禹辰, 勾智楠, 陈宇欣, & 秦乐. (2025). 基于大语言模型的多任务生成式重构对话情绪识别. *Application Research of Computers/Jisuanji Yingyong Yanjiu*, *42*(7).

D. 劉廷緯. (2024). 更高效的語音處理: 低資源情境下的自監督式學習. *臺灣大學電信工程學研究所學位論文*, 1-179.

E. 陳元瑞. (2020). 藉助跨語言聲音單位對映之遷移學習達成使用低資源之端到端語音合成及辨識. *國立臺灣大學資訊工程學系學位論文*, 1-77

F. Hsieh, Y. C., Lyu, K. M., & Lyu, R. Y. (2023). 運用基於生成預訓練轉換器架構的 OpenAI Whisper 多語言語音辨識引擎之台語及華語語音辨識之實作. In *35th Conference on Computational Linguistics and Speech Processing, ROCLING 2023* (pp. 210-214). The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

G. Wang, S., Yang, C. H., Wu, J., & Zhang, C. (2024, April). Can whisper perform speech-based in-context learning?. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 13421-13425). IEEE.

H. 江宥呈. (2023). 中低資源語言之語音語料蒐集及語言辨識之分析研究. *國立臺灣大學電機工程學系學位論文*, 1-64.

I. Haxhibeqiri, J., De Poorter, E., Moerman, I., & Hoebeke, J. (2018). A survey of LoRaWAN for IoT: From technology to application. Sensors, 18(11), 3995.

J. Vangelista, L. (2017). Frequency shift chirp modulation: The LoRa modulation. *IEEE signal processing letters*, *24*(12), 1818-1821.

Whisper 微調與 Branchformer 於客語語音辨識之應用
Applying Whisper Fine-tuning and Branchformer to Hakka Speech Recognition

Yu-Sheng Huang
National Ilan University
peilun2016@gmail.com

Wei-Cheng Hong
National Ilan University
q0989323897@gmail.com

Xin-Yu Chen
National Ilan University
cxinyu153@gmail.com

Szu-Yin Lin
National Yang Ming Chiao Tung University
szuyinlin@gmail.com

摘要

本研究針對 FSR 2025 客語辨識任務，比較大型預訓練模型微調與從頭訓練兩種策略。漢字辨識部分，透過微調五種不同規模的 Whisper 模型，large-v3-turbo 在測試集達到 7.55% CER。拼音辨識部分，則比較 Branchformer 與採用 LoRA 微調的 Whisper-small，兩者在測試集的 WER 分別為 4.7%與 6.5%。在資料前處理方面主要採用速度擾動進行資料增強。

Abstract

This study addresses the FSR 2025 Hakka speech recognition task by comparing two strategies: fine-tuning large pre-trained models and training from scratch. For character (Hanzi) recognition, we fine-tuned five different scales of the Whisper model, with large-v3-turbo achieving a 7.55% CER on the test set. For Pinyin recognition, a Branchformer model was compared against a LoRA fine-tuned Whisper-small, yielding WERs of 4.7% and 6.5% on the test set, respectively. Speed perturbation was the primary method used for data augmentation in our pre-processing pipeline.

關鍵字：客語、ASR、Whisper、Branchformer
Keywords: Hakka, ASR, Whisper, Branchformer

1 Introduction

自動語音辨識(ASR)技術近年來因深度學習模型的突破而快速發展，端到端(End-to-End, E2E)模型與大型預訓練語音模型已成為主流。Formosa Speech Recognition Challenge (FSR)的主要任務是客語語音辨識，對於瀕危語言的保存具有重要意義。

回顧 2023 年的 FSR，比賽團隊針對客語 ASR 採用了多種方法。在模型架構上，E2E 模型(如 Conformer、Branchformer、Zipformer transducer 及 Hybrid CTC/Attention)被廣泛應用，以捕捉語音的時序與長距離依賴(Chang and Chen, 2023; Lu et al., 2023a; Su et al., 2023)。Whisper 與 WavLM 等大型預訓練模型也被廣泛使用(Lu et al., 2023a; Chiang et al., 2023; Huang and Tsai, 2023)，透過少量客語資料結合參數高效微調(PEFT，如 LoRA、AdaLoRA)，可達到良好的辨識效果。同時 Wav2vec2.0 與 HuBERT 等自監督學習(SSL)模型常用作前端特徵提取器，從未標記語音中學習表示(Hu and Chen, 2023; Yang et al., 2023)。

為解決資料稀缺與雜訊問題，多型態訓練(MTR)、頻譜擴增(SpecAugment)、速度擾動(Speed Perturbation)等資料擴增技術被廣泛使用，以提升模型穩健性(Chang and Chen, 2023; Yang et al., 2023; Lu et al., 2023b)。部分系統結合淺層融合(Shallow Fusion)、N-best Rescoring 等後處理方法，利用額外文本語料改善辨識結果，也有研究採用基於 BERT 的 pBERT 重新計分(Lu et al., 2023a; Yang et al., 2023)以及使用語音活性檢測(Voice Activity Detection, VAD)則用於去除靜音片段，提升辨識效率(Chen et al., 2023)。

儘管過往的比賽已經取得多方面進展，但仍存在挑戰，例如如何進一步提升模型的多樣性、縮小訓練與測試資料之間的差異，以及處理客語羅馬拼音與漢字轉換的問題。今年的比賽中，我們將延續過往的經驗與成果，探索更適合客語語音辨識的解決方案。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 441–445
November 20-22, 2025 ©2025 Association for Computational Linguistics

2 Methods

2.1 ESPnet

ESPnet(End-to-EndSpeechProcessingToolkit)，是一個開源、端到端的語音處理工具包，主要基於 PyTorch 深度學習框架進行開發，並延續 Kaldi 風格的資料處理流程。而後來推出的 ESPnet2 採用 YAML 設定 + recipe 的模組化設計，易於重現實驗與切換架構。ESPnet2 的核心優勢之一是其模型庫的豐富性與靈活性，使用者可以在設定檔中輕鬆切換和配置不同的後端模型架構，例如：Transformer、Conformer、或是基於 RNN 的經典模型等。ESPnet2 針對不同的應用任務提供了不同方法，例如：自動語音辨識(ASR)、文字轉語音(TTS)、語音增強/分離(SE/SS)、語音翻譯(ST)與口語理解(SLU)等主流領域。除了內建的核心模型，ESPnet 的框架還支援最新的大型預訓練模型(如 Whisper)進行整合與支援 LoRA 微調。

2.2 Whisper

Whisper 是由 OpenAI 所開發的一套開源自動語音辨識(ASR)系統(Radford et al., 2022)，該系統經過 680,000 小時的訓練，使用多語言及多任務的監督式資料，提升系統在口音、背景噪音及技術性語言上的穩定性。且支援多國語言的語言辨識。

Whisper 模型屬於典型的 Transformer 架構(Vaswani et al., 2023)，採用 Encoder-Decoder 的 Attention 機制。模型前端包含兩層一維卷積層(濾波器大小為 3，啟用函數為 GELU)，第二層卷積的步長為 2，用於對輸入的梅爾頻譜特徵進行下採樣。輸入音訊會重採樣至 16kHz，並計算 80 維 log-magnitude Mel spectrogram，視窗大小 25 毫秒、步長 10 毫秒。處理後的數值正規化至[-1, 1]，並近似零均值。以 30 秒音訊片段為例，可得到 3000×80 的特徵矩陣，經兩層卷積後縮減為 1500×80。卷積輸出再加入位置編碼，其中編碼器使用 sinusoidal positional embedding，解碼器則使用 learned positional embedding。Transformer 區塊採用 pre-activation residual blocks(Child et al., 2019)，並於編碼器輸出端施加最終層正規化。解碼器則使用 tied input-output embeddings(Press and Wolf, 2017)。

標記器部分採用基於 GPT-2 的 Byte-Pair Encoding (BPE)，包括 tiny、base、small、medium 與 large，其中 large 分為 large-v1、large-v2 和 large-v3，且 large-v3 的整體表現最佳。針對不同語言，英文部分會直接沿用 GPT-2BPE，其餘語言的模型將重新擬合詞彙分布，但保持詞表大小不變，以避免非英文語言的過度斷詞。Whisper 的不同模型版本依層數與參數量劃分：最小的 tiny 模型包含 4 層 encoder 與 4 層 decoder，參數量為 39M；最大的 large 模型包含 32 層 encoder 與 32 層 decoder，參數量為 1550M。除了直接使用官方釋出的模型外，亦常透過微調(fine-tuning)的方式，針對特定領域或語言進行再訓練，以進一步提升模型在專業場景中的辨識效能與適應性。

3 Experiments

3.1 資料集

我們使用 Formosa Speech Recognition Challenge 2025 - Hakka ASR II 競賽官方提供的訓練資料集 FSR-2025-Hakka-train，以及熱身賽資料集 FSR-2025-Hakka-evaluation 為音檔資料集與熱身賽資料集 FSR-2025-Hakka-evaluation-key 為熱身賽資料集的標準答案。而資料集分割的部分我們將訓練資料集隨機打亂後依照 8:1:1 的比例各切成 train、dev、test 三份。訓練資料集的切分(見表 1)。

資料集切分	句數	時長(hours)
train	21879	49.64
dev	2735	6.17
test	2735	6.19

表 1. 訓練資料集切分比例

整理熱身賽資料集時，我們發現兩者以話語 ID 對齊後，FSR-2025-Hakka-evaluation 比 evaluation-key 多出 105 筆音檔而無答案。因此僅在 FSR-2025-Hakka-evaluation 與 FSR-2025-Hakka-evaluation-key 的交集 4,299 筆上計算，熱身賽資料集的配置(見表 2)。

	句數
FSR-2025-Hakka-evaluation-key	4299
FSR-2025-Hakka-evaluation	4404

表 2. 熱身賽資料集的配置

3.2 評估方式

模型評估採用字元錯誤率(Character Error Rate, CER)、詞錯誤率(Word Error Rate, WER)與句子錯誤率(Sentence Error Rate, SER)。CER 表示語音轉寫內容在字元層級的準確性，能夠反映模型對於每個字的辨識效果，用於衡量模型在實際應用上的轉寫精確度。WER 則為詞錯誤率(Word Error Rate, WER)，以詞為單位，計算預測結果相較於正確答案所發生的替代、刪除與插入錯誤總數來衡量轉寫的準確性。而 SER 是計算模型整句預測是否與正解完全一致，若句子中有任一字元錯誤則判為錯誤，為更嚴格的指標，評估在完整語句層級的表現。

3.3 客語漢字

本節比較 Whisper 系列多個規模的模型在客語語音辨識任務中的微調表現，所使用之模型包含 tiny、base、small、medium 與 large-v3-turbo。

模型訓練使用 HuggingFace Transformers 與 PyTorch 架構實作，並整合 Whisper 提供之 WhisperProcessor，進行特徵擷取、分詞與標註等前處理工作。主要訓練參數設定如下：音訊採樣率為 16kHz，最大音訊長度限制為 30 秒，文字最大生成長度為 128。訓練採用 mini-batch 大小 16，並進行 4 次梯度累積以模擬較大批次；學習率設定為 1e-5，總訓練週期為 10epochs。在模型選擇上，以驗證集 CER 分數最低之 checkpoint 作為最佳模型並保存。

3.4 客語漢字實驗結果

使用 Whisper 不同規模的模型進行客語漢字辨識實驗，以下整理各模型在測試集上的漢字辨識結果(見表 3)

Dataset	Model	CER	SER
測試集	tiny	28.37	82.78
	base	15.28	67.71
	small	11.89	57.26
	medium	39.41	56.86
	large-v3-turbo	7.55	33.81

表 3. 各模型在測試集中的漢語辨識結果

從測試集的結果可觀察到，隨著模型規模增大，CER 整體呈現下降趨勢，其中 large-v3-turbo 在測試集上取得最佳表現，CER 為 7.55%，SER 也降至 33.81%，顯示其對客語漢字辨識具有較佳能力。

在比賽提供的熱身賽資料集上的漢字辨識表現結果(見表 4)。

Dataset	Model	CER	SER
熱身賽資料集	tiny	33.18	84.28
	base	25.76	73.71
	small	22.27	65.22
	medium	13.61	49.85
	large-v3-turbo	22.78	66.18

表 4. 各模型在熱身賽資料集中的漢語辨識結果

在比賽提供的熱身賽資料集中，表現最佳的反而是 medium 模型，其 CER 為 13.61%，SER 為 49.85%；而 large-v3-turbo 在該資料集的 CER 為 22.78%，表現不如預期。

根據比賽官方網站熱身賽的說明，baseline 採用的模型為 large-v3-turbo。儘管本研究針對該模型進行微調，但在熱身賽資料集上的 CER 表現仍未能超越 baseline 的 10.42%，顯示模型在特定資料上辨識能力仍有提升的空間。

造成 CER 上升的可能因素包括語者差異、背景噪音干擾、語速變化及資料分布不均等問題。

3.5 客語拼音

本研究在客語拼音部分比較兩種基於 ESPnet 所實現模型，分別為從頭訓練的 Branchformer(CTC/Attention)混合訓練，以及對大型預訓練模型 Whisper 進行參數高效微調。兩種方法皆採用了速度擾動 (Speed Perturbation)作為共通的資料增強手段，將訓練語音以 0.9、1.0 及 1.1 三種不同語速進行資料增強。

首先我們嘗試的方法為 Branchformer (CTC/Attention)混合訓練，模型前端將原始音訊轉換為 FBANK 聲學特徵，並在頻譜圖上應用 SpecAugment 進行進一步的資料增強。模型的核心架構由一個包含 12 個區塊的 Branchformer 編碼器與一個包含 6 個區塊的 Transformer 解碼器所組成。訓練過程中，採用混合式 CTC/Attention 的訓練方法，將 CTC 損失與注意力導向的交叉熵損失進行加權，並使用標籤平滑作為正規化手段。優化器選用 Adam，搭配 WarmupLR 學習率。在解碼階

段，系統使用寬度為 20 的波束搜尋，並將 CTC 分數與解碼器分數進行聯合解碼。為了進一步提升辨識準確率，額外訓練並使用一個基於 4 層 Transformer 的 BPE-300 子詞級語言模型。該語言模型在解碼階段透過二次評分的方式被整合進系統，其權重被設定為 1.0，以優化最終輸出的語法與流暢度。

第二種方法採用參數高效微調(PEFT)策略。與方法一不同，在前端直接以原始音訊波形作為輸入。模型的核心是使用 OpenAI 的預訓練模型 Whisper small 版本作為基礎編碼器與解碼器。為了更好的輸出效果，我們採用了 LoRA(Low-Rank Adaptation)技術進行微調。凍結 Whisper 模型的原始權重，僅在 Transformer 注意力機制的 query, key, value 層中注入可訓練的低秩矩陣。模型的訓練目標僅為注意力導向交叉熵損失。在解碼階段，系統採用寬度為 10 的波束搜尋。

3.6 客語拼音實驗結果

漢字拼音部分，我們嘗試三種模型組合作為比較，以下整理各方法在測試集上的漢字拼音辨識結果(見表 5)

Dataset	Model	WER	SER
測試集	BRF	4.9	38.5
	BRF+ LM	4.7	38.7
	WSP_SM + LoRA	6.5	37.4

表 5. 各方法在測試集中的拼音辨識結果

其中 Branchformer(BRF, CTC/Attention)方法的表現為 WER 4.9%、SER 38.5%；在 BRF 上加入 Transformer 的 BPE-300 子詞級語言模型進行二次重評分(BRF+LM)後，WER 下降至 4.7%，但 SER 略升至 38.7%，顯示 LM 有助於 WER 修正，對整句完全正確的比例未必同步提升。Whisper small 採用 LoRA 微調(WSP_SM+LoRA)其表現 WER 為 6.5%，高於 BRF 系列，但 SER 為 37.4%，為三者最佳，反映其較強的語言模型能提高句子完整度。

接著為各方法在比賽提供的熱身賽資料集上的辨識拼音辨識表現結果(見表 6)。

Dataset	Model	WER	SER
熱身賽資料集	baseline		23.4
	BRF	30.3	71.7
	BRF+ LM	54.4	99.0
	WSP_SM + LoRA	35.0	58.6

表 6. 各方法在熱身賽資料集中的拼音辨識結果

在熱身賽資料集中，官方 baseline 表現最佳，SER 為 23.4。而 BRF 模型訓練後，WER 為 30.3、SER 71.7，整體落後 baseline。進一步加入外部語言模型(BRF + LM)後，WER 升至 54.4、SER 幾近 99.0，我們認為可能是 LM 權重設定過高所致，導致解碼分數被 LM 主導而產生錯誤預測。相較之下，採用 Whisper small + LoRA 的參數高效微調，WER 35.0、SER 58.6，雖未優於 baseline，但 SER 顯著低於 BRF 的 71.7，顯示大型預訓練模型的穩定度較佳。

4 Conclusion

本研究在漢字辨識部分，我們微調不同規模 Whisper 模型。在自行切分的測試集中，large-v3-turbo 模型憑著模型參數規模的優勢，取得 7.55%的最佳字元錯誤率(CER)。在拼音辨識部分，我們比較 Branchformer 模型與基於 Whisper small 進行 LoRA 微調的模型。實驗結果顯示 Branchformer(CTC/Attention)在測試集上的詞錯誤率(WER)為 4.9%，加入外部語言模型後進一步下降至 4.7%；相較之下，Whisper small + LoRA 的 WER 為 6.5%，雖然略高於 Branchformer，但在句子錯誤率(SER)上則達到 37.4%，比 Branchformer 的 38.5%還低，其在句子完整度上更具優勢。

在未來研究我們認為可朝以下方向進行改善：加入背景噪音處理機制，提升模型在實際環境音下的辨識穩定性；可引入資料增強技術，如聲音混合、頻譜遮蔽(SpecAugment)等，提升模型對聲音變異的適應能力。也可針對腔調進行分流訓練或採用語者標註進行語者自適應，提升在不同腔調與跨語者場景下的表現一致性。透過上述策略，有望提升模型泛化能力，提高模型在真實應用場景中的實用性與準確率。

References

Hsiu-Jui Chang and Wei-Yuan Chen. 2023. The DMS-ASR System for the Formosa Speech Recognition Challenge 2023. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 377–

379, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Po-Kai Chen, Bing-Jhih Huang, Chi-Tao Chen, Hsin-Min Wang, and Jia-Ching Wang. 2023. Enhancing Automatic Speech Recognition Performance Through Multi-Speaker Text-to-Speech. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 371–376, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Ming-Hsiu Chiang, Chien-Hung Lai, and Hsuan-Sheng Chiu. 2023. WhisperHakka: A Hybrid Architecture Speech Recognition System for Low-Resource Taiwanese Hakka. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 390–396, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating Long Sequences with Sparse Transformers. arXiv:1904.10509 [cs].

Hong-Jie Hu and Chia-Ping Chen. 2023. NSYSU-MITLab Speech Recognition System for Formosa Speech Recognition Challenge 2023. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 380–385, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yi-Chin Huang and Ji-Qian Tsai. 2023. Whisper Model Adaptation for FSR-2023 Hakka Speech Recognition Challenge. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 423–427, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Hao-Chien Lu, Chung-Chun Wang, Jhen-Ke Lin, and Tien-Hong Lo. 2023a. The NTNU ASR System for Formosa Speech Recognition Challenge 2023. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 397–402, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yuan-Hsiang Lu, Chung-Yi Li, and Zih-Wei Lin. 2023b. The Taiwan AI Labs Hakka ASR System for Formosa Speech Recognition Challenge 2023. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 403–408, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Ofir Press and Lior Wolf. 2017. Using the Output Embedding to Improve Language Models. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [eess].

Jia-Jyu Su, Dong-Min Li, and Chen-Yu Chiang. 2023. A preliminary study on Hakka speech recognition by using the Branchformer. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 409–413, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs].

Tzu-Ting Yang, Hsin-Wei Wang, Meng-Ting Tsai, and Berlin Chen. 2023. The NTNU Super Monster Team (SPMT) system for the Formosa Speech Recognition Challenge 2023 - Hakka ASR. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 414–422, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Improving Low-Resource Speech Recognition with Whisper-MoE and Synthetic Data Augmentation: A Case Study on Hakka
基於 Whisper-MoE 與合成資料增強的低資源語音辨識改進研究：以客家話為例

Yuan-Chi Hsu
National Kaohsiung University of Science and Technology
Department of Electrical Engineering
F113154165@nkust.edu.tw

Liang-Chun Fang
National Kaohsiung University of Science and Technology
Department of Electrical Engineering
F113154169@nkust.edu.tw

Hong-Jie Dai
National Kaohsiung University of Science and Technology
Department of Electrical Engineering
hjdai@nkust.edu.tw

摘要

本研究其目的於如何提高在低資源特定種族的客家語音辨識能力，本團隊透過微調不同基底的 Whisper（如原生與針對中文微調的 Belle 模型）的方式進行實驗，我們發現在客家文字和拼音的任務中微調不同基底的模型會有各自較好且不同的實驗結果。本團隊為了更進一步提升模型的準確率，實驗了在 Whisper Encoder 的 attention 層中的 q、k、v 線性層替換成混和專家模型結合 RoLA，及合成的語音有來自不同風格的聲音以及不同的講話速度，結果顯示任務一中字元錯誤率下降了 0.73%，任務二中字詞錯誤率下降了 0.02 %。可以從結果確認說調整模型架構與在少量的方言語料中有策略地使用少量合成語音是可以提升模型的辨識能力。

Abstract

The objective of this study is to improve speech recognition performance for low-resource Hakka, a language spoken by a specific ethnic group. Our team conducted experiments by fine-tuning different base versions of Whisper (e.g., the original model and the Mandarin-focused Belle model). We found that fine-tuning on different bases yielded distinct advantages and varying results in Hakka character and phonetic recognition tasks. To further enhance model accuracy, we experimented with replacing the q, k, and v linear layers in the attention blocks of the Whisper encoder with a mixture-of-experts model combined with RoLA. In addition, we augmented the training data with synthesized speech generated with diverse voice styles and varying speaking rates. The results showed a 0.73% reduction in character error rate for Task 1 and a 0.2% reduction in word error rate for Task 2. These findings confirm that both architectural adjustments to the model and the strategic use of limited synthetic speech data in low-resource dialect corpora can effectively improve recognition performance.

關鍵字：客家語音辨識、Whisper、Data Augmentation
Keywords: Hakka ASR, Whisper, Data Augmentation

1 Introduction

近幾年隨著算力資源的進步，語音辨識模型一直不斷地創新及突破，在主要的語言如英文、中文、法文等都有著非常好的表現。但台灣屬於一個多元文化的國家，光是方言就有國語、台語、客語還有各種不同的原住民語。但隨著時代演變．在生活中基本上越來越少出現這些方言，因此有相關研究希望能藉由收集語音資料並開發相關語言模型讓這些文化得以延續。隨著 seq2seq 架構的推出，與後來的 Attention (Vaswani, Shazeer et al. 2017) 機制讓模型可以去考慮更多上下文，在這樣的條件下產生的 Transformer 架構，廣泛應用在機器翻譯、語音辨識、圖像識別。本研究是使用當前在語音辨識（Automatic Speech Recognition，簡稱 ASR）主流的模型 Whisper (Radford, Kim et al. 2023)，透過大量資料、不

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 446–449
November 20-22, 2025

同語音任務像 ASR、Speech Translation 及單一網路架構，使得 Whisper 可以執行多種語言的任務如語音辨識、語音翻譯。在有著基礎的預訓練模型的基礎之上，透過 Formosa Speech Recognition Challenge 2025 的資料集微調成可以辨識大埔腔、詔安腔的客家語音辨識模型。

本研究還運用了多種策略去提升模型的性能，在任務上採用修改原始 whisper 架構新增 MoE-RoLA 讓模型在混雜的情況下，可以更正確的判斷語音的正確腔調，並使用對應的專家模型進行語音辨識。另外為了可以更好的提升模型的性能，我們使用了 VoxHakka(Chen, Lee et al. 2024) TTS 系統 進行語音資料的擴增，合成的文本內容來自教育部臺灣客語辭典(教育部) 之詞彙，語音的風格多變，避免模型發生過擬和。

2 Approach

2.1 Whisper

ASR 領域中 Whisper 的表現亮眼，其重要的原因是他經過 68 萬小時的標註資料，進行了監督式學習，在英文的表現能力上達到跟人類一樣的能力。此舉證明了透過學習大量且多元的資料，可以提升模型在對於口音、噪音的適應性。另外 Whisper 有一個特點就是他擁有強大的適應力，因為本身有著強健的英語能力，透過遷移式的方式使其對其他語言也能夠熟練，甚至是方言(Chen, Huang et al. 2023)。

在 Whisper 中也有嘗試其他微調後的模型進行二次微調其中包含 LianjiaTech 開發的中文語音模型 Belle-whisper-larger-v3-zh(以下簡稱 Belle)，該模型是基於 Whisper-large-v3 進行微調，透過在語音頻譜上的時間軸上分別隨機做 mask 及透過模擬的方式增加噪音使辨識及泛化性能力能夠進一步的提升。

2.2 Mixture of Expert

Mixture-of-Experts (MoE) (Shazeer, Mirhoseini et al. 2017)是一種提升模型任務多元性的架構，透過引入多個專家並由 gating network 動態選擇部分專家參與計算。這種設計能在推理成本近似固定的情況下，大幅增加模型的參數規模與表達能力。在語音處理中，MoE 特別適合多語言與多方言場景，因為不同專家可以捕捉不同語言或腔調的特徵，而 gating network 則能根據輸入自適應分配最合適的專家，從而改善低資源語言的辨識效果。

2.3 MoE-RoLA

MoE-RoLA 是將 MoE 與 Rank-One Low-rank Adaptation (以下簡稱 RoLA) (Hu, Shen et al. 2022) 結合的一種參數高效化方法，架構如圖 1。其核心概念是在預訓練模型的特定層數（如自注意力投影層）中，保留原始權重不變，並在其上引入多個低秩增量專家 (RoLA experts)。每個專家僅包含極少量可訓練參數，而 gating network 根據輸入特徵動態選擇或加權專家輸出。這種設計同時具備 MoE 的專家分工能力 與 RoLA 的參數高效特性：MoE 機制允許不同專家專注於不同語言、方言或任務條件，而 RoLA 保證每個專家增量極小，大幅降低微調成本。透過 MoE-RoLA，模型能在保持參數高效的前提下，動態適應輸入特性，特別適合低資源或多樣化的語音場景。

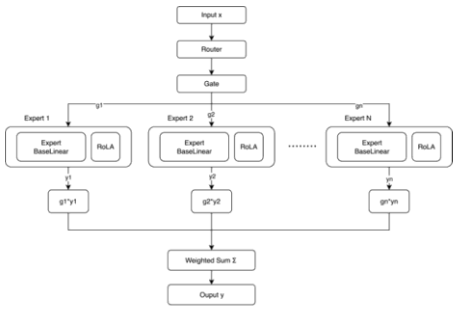


圖 1、MoERoLA 架構

2.4 VoxHakka 文字轉語音系統

VoxHakka 是一個專為臺灣客家語音文字轉語音（Text-To-Speech，簡稱 TTS）的系統，可進行 6 種腔調的合成: 四縣、海陸、大埔、饒平、詔安與南四縣。此系統基礎是 YourTTS(Casanova, Weber et al. 2022)框架，此框架的特點是可針對多語言、多說話者進行 TTS 的開發，VITS(Kim, Kong et al. 2021)的架構使我們可以將文字轉換為高品質、自然的語音。由於 VoxHakka 缺乏開源的資料，對於資料的蒐集提出一種有效率的策略，利用網路爬蟲結合 ASR 資料清理技術，確保建立的資料集品質。由於現階段的臺灣客家語音合成，沒有四縣腔外的公開 TTS 系統可進行比較，因此測試上採用比較平均意見分數（CMOS）評估聽眾在自然度、發音準確、聲調正確性。VoxHakka 系統在自然度方面表現優於其他模型，聲調正確性上接近人類。

3 Experiments

本研究所使用的硬體規格如 Table 1 所示。

CPU	i7-13500
GPU	RTX4090 24G
RAM	128G

Table 1、電腦規格

3.1 Data and Model

3.1.1 Dataset

在本研究中我們的資料集來自以下:

HAT-Vol2: 此資料由比賽主辦方所提供，包含了 100 個語者總時長 80 小時的資料，包含了 27,349 個樣本的訓練集與 3,458 個樣本驗證集。並將資料集分割為 8:2 的訓練資料與驗證資料。

Generated: 透過使用 Huggingface 提供的 TTS API 生成資料，文本內容為教育部臺灣客語辭典提供的大埔腔、詔安腔詞目，將文字內容轉換成語音，從而生成了約 5 小時的語音資料。

3.1.2 Model Configuration

模型配置上，由於兩者任務目標的語系不同，故在模型選擇上有差異。對於客家文字任務，我們使用 Huggingface 平台上的 BELLE-2/Belle-whisper-large-v3-zh，而對於客家拼音任務我們使用 whisper/medium，以上模型皆進行全參數微調，以及MoERoLA 的模型調整。拼音任務有使用資料增強的方式進行微調。Table 2 為 MoERoLA 的主要配置，EXPERTS 為 MOE 所生成的專家數量，該專家數量不一定要等同於語系數量，多出的專家可以學習更細緻的分工（例如不同說話人、口音、噪音條件、語速），RANK 決定增量權重的複雜度（RoLA 固定為 1，僅捕捉一個方向），而 ALPHA 則控制這個增量在最終模型中的影響力大小。

超參數	文字 \ 拼音任務
EXPERTS	6 \ 2
RANK	16 \ 32
ALPHA	32 \ 64
DROPOUT	0.05

Table 2、MoERoLA 參數設定

3.1.3 Fine-tuning detail

本研究在字元任務及拼音任務方面為獨立實驗，但使用方法及依賴類似的，皆以 Dataset 作為資料集型態，訓練模型以 Hugging Face Transformers 套件為基礎進行調整。Table 3 顯示字元任務中訓練使用的超參數設定，包含了 Batch、Early Stop、混合精度、學習率、餘弦退火及優化器等設置，Table 4 則顯示了在拼音任務中訓練使用的超參數。

超參數	Value
Batch	4
學習率	1e-5
優化器	adamw_bnb_8bit
EarlyStop	5
混合精度	FP16
餘弦退火	500

Table 3、字元任務訓練超參數設定

超參數	Value
Batch	8
grad Accum Steps	2
學習率	2.5e-4
優化器	adamw_bnb_8bit
Early Stop	5
混合精度	FP16
餘弦退火	500

Table 4、拼音任務訓練超參數設定

3.2 Evaluation of Character Track

在字元實驗中研究了在純微調模型及嵌入MoERoLA 調整各項參數後分數的變化如 Table 5，可發現 Expert 的參數調整對與模型的分數變化並沒有太大的影響，但其中發現在驗證集方面原本在內部測試成績 CER 為 2.76 %，但在官方的熱身賽測試集 CER 結果卻為 13.36 %，表示該模型對於不同類型的資料泛化性不足，希望藉由調整模型架構可以加強整體模型的效能。

Model	CER
Belle	2.70 %
Belle+ 2 Expert	**1.97 %**
Belle+ 6 Expert	2.27 %

Table 5、字元任務驗證集測試分數

3.3 Evaluation of Pinyin Track

在拼音任務中將僅透過 HAT-Vol2 微調的 whisper/medium 作為我們的 baseline，Table 6 展示在拼音任務 whisper 微調與兩個應用策略後的分數變化可以觀查到 WER 達到 7.22%。緊接著為了讓模型可以學到更多的客家語特徵，我們透過合併了 HAT-Vol2 與 TTS 合成的語音資料進行微調，但效能反而下降非常嚴重，推測原因是合成語料與原先資料差異太大，導致模型無法正確學習。最終嵌入MoERoLA 的 whisper/medium 效果會比 baseline 好，WER 下降了 0.2%。

Model	WER
Whisper/medium	7.22 %
Whisper/medium + Data Augmentation	36.13%
Whisper/medium + 2 Expert	**7.20 %**

Table 7、拼音任務驗證集測試分數

4 Conclusion

本次研究透過 whisper ASR 架構去進行客家語音辨識系統的開發，實驗結果表明此架構僅透過微調即可在新語言資料集上達到非常好的效果。由於熱身賽與決賽錄音中可以明顯感覺有較多環境音與不同風格的內容，因此若無法開發具有泛化能力強的模型即無法再比賽拿到高分，因此未來我們希望可以再去鑽研更多不同的資料資強技巧，尤其是在增加噪音特徵到原始資料的情況，去增強模型的性能，使模型在實際場域中可以達到實驗時一樣的效能。總而言之，我們採取的混和專家策略確實提升了模型的效能。此外透過外部擴增的資料集無法有效提升性能，推測原因是內容與目標資料差異太大。在驗證集上客家文字 CER 降低了 0.73%，客家拼音 WER 降低了 0.02%。結果表明透過專家系統可以有效提升混和不同腔調的客家語音辨識模型性能。

References

Casanova, E., et al. (2022). Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. International conference on machine learning, PMLR.

Chen, L.-W., et al. (2024). VoxHakka: A Dialectally Diverse Multi-Speaker Text-to-Speech System for Taiwanese Hakka. 2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE.

Chen, P.-K., et al. (2023). Enhancing Automatic Speech Recognition Performance Through Multi-Speaker Text-to-Speech. Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023).

Hu, E. J., et al. (2022). "Lora: Low-rank adaptation of large language models." ICLR **1**(2): 3.

Kim, J., et al. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. International Conference on Machine Learning, PMLR.

Radford, A., et al. (2023). Robust speech recognition via large-scale weak supervision. International conference on machine learning, PMLR.

Shazeer, N., et al. (2017). "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." arXiv preprint arXiv:1701.06538.

Vaswani, A., et al. (2017). "Attention is all you need." Advances in neural information processing systems **30**.

教育部. "台灣客語辭典." from https://hakkadict.moe.edu.tw/.

Whisper Finetuning For Hakka Recognition in Low Resource

Min Han Teng[1] Ci Dao Chen[1] You Ting Lin[1] Bing Jhih Huang[1] Jia-Ching Wang[1]
[1]Department of Computer Science, National Central University, Taiwan
olivier40103@gmail.com, copeman123@gmail.com, eeyore0624@gmail.com, c725992@gmail.com

Abstract

We study automatic speech recognition (ASR) for Hakka, a low-resource language with substantial dialectal variation. Focusing on Zhaoan and Dapu, we fine-tune Whisper using Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning (AdaLoRA) and apply data augmentation to mitigate data scarcity. Experiments show that AdaLoRA combined with augmentation substantially improves cross-dialect recognition while maintaining parameter efficiency. Our results demonstrate the potential of lightweight adaptation to extend large-scale ASR systems to underrepresented languages, supporting the preservation of Hakka speech and orthography.

1 Introduction

Large-scale pretrained automatic speech recognition (ASR) models, such as Whisper, have achieved strong performance on high-resource languages. However, their applicability to low-resource and dialectally diverse languages remains underexplored. Hakka, a Sinitic language with millions of speakers worldwide, is particularly underrepresented in ASR research despite its cultural and linguistic significance. The lack of standardized resources, limited digital presence, and substantial phonological variation across dialects pose major challenges for building robust ASR systems and heighten the risk of language endangerment.

Among the many Hakka dialects, Zhaoan and Dapu represent two major varieties whose systematic phonological differences further complicate recognition. Developing ASR systems that generalize across these dialects is especially difficult under low-resource conditions.

To address these challenges, we fine-tune Whisper for Hakka using Low-Rank Adaptation (AdaLoRA), a parameter-efficient method well-suited for low-resource adaptation of large-scale pretrained models. To further mitigate data scarcity and improve cross-dialect robustness, we incorporate data augmentation techniques that expand the training set and enhance generalization.

This work not only provides a practical solution for advancing Hakka ASR, but also offers a replicable framework for extending large-scale ASR models to other underrepresented languages, contributing to both language preservation and applied speech technologies.

Our main contributions are as follows:

- We propose training strategies and data augmentation techniques that improve model performance and robustness in low-resource, multi-dialect settings.
- We empirically demonstrate consistent gains in cross-dialect recognition, measured by pinyin WER and character CER, validating the effectiveness of lightweight adaptation for underrepresented languages.

2 Related Works

Automatic Speech Recognition (ASR) Large-scale pretrained ASR models have greatly advanced multilingual speech recognition. Whisper (Radford et al., 2022),achieves state-of-the-art results in high-resource languages but struggles in low-resource settings due to limited data and dialectal variation. Prior work has explored transfer learning (Wang et al., 2021), and parameter-efficient fine-tuning (Zhang et al., 2023; Liu et al., 2022). Low-Rank Adaptation (AdaLoRA) (Zhang et al., 2023) has proven particularly effective for adapting large ASR models, making it suitable for extending Whisper large-v2 to Hakka.

Text-to-Speech (TTS) TTS has progressed with multilingual and multispeaker pretrained models. Neural approaches such as Tacotron 2 (Shen et al., 2018) and FastSpeech (Ren et al., 2019) achieved natural synthesis in high-resource languages, while

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 450–453
November 20-22, 2025 ©2025 Association for Computational Linguistics

YourTTS (Casanova et al., 2022), based on VITS (Kim et al., 2021), introduced zero-shot multilingual, multispeaker capabilities. Trained on VCTK (Yamagishi et al., 2019), YourTTS enables speaker adaptation with under one minute of speech. Beyond synthesis, TTS has been used to generate synthetic data for ASR in low-resource settings (Rosenberg et al., 2019; Li et al., 2020). Here, we adopt YourTTS to augment scarce Hakka corpora for more robust ASR training.

Data Augmentation for Speech Data augmentation is widely used to improve ASR robustness in low-resource contexts. Common methods include noise injection, Utterance concatenation, Time stretching, Pitch shifting, Air absorption, and Environmental impulse response (Ko et al., 2017; Zahid and Qazi, 2025; van der Meer, 2022; Kates and Brandewie, 2020; Bryan, 2019). More recent work explores cross-lingual transfer and TTS-based generation (Hsu et al., 2020; Rosenberg et al., 2019). SpecAugment (Park et al., 2019) is now standard in ASR, while TTS-based augmentation (Li et al., 2020; Jia et al., 2019) creates labeled data for low-resource languages.

3 Method

3.1 ASR Model Fine-tuning

We adopt the Whisper large-v2 model (Radford et al., 2022) as the backbone for Automatic Speech Recognition (ASR). To efficiently adapt the large-scale model to underrepresented Hakka dialects, we employ adaptive Low-Rank Adaptation (AdaLoRA) (Zhang et al., 2023), which enables parameter-efficient fine-tuning without retraining the full model. This setup allows the model to retain general multilingual knowledge while specializing in Hakka recognition.

3.2 Synthetic Data Generation with TTS

To address the scarcity of annotated Hakka speech, we leverage YourTTS (Casanova et al., 2022), a multilingual and multispeaker zero-shot TTS system based on the VITS architecture (Kim et al., 2021). We use YourTTS to synthesize Hakka utterances across different accents, thereby enriching the training corpus and improving model generalization. In addition, we specifically generate synthetic data for coarticulated syllables using self-collected and organizer-provided texts, ensuring better phonological coverage. Our training corpus combines both human-recorded and synthetic speech. The human data consist of Hakka Dapu and Zhaoan recordings provided by the competition organizers. To complement this limited corpus, we generated synthetic speech using YourTTS across multiple accents. In addition, we produced accent-specific utterances focusing on coarticulated syllables to improve phonological coverage and enhance recognition robustness.

3.3 Data Preprocessing and Augmentation

To further improve robustness under low-resource and cross-dialect conditions, we applied a set of augmentation techniques during training, including noise injection, utterance concatenation, time stretching, pitch shifting, air absorption, and environmental impulse response. These methods enrich the acoustic diversity of the training data, enabling the model to better recognize Hakka speech across varied speaking styles and noisy environments.

4 Experiments

4.1 Experimental Setup

Dataset

The dataset comprises both human-recorded and synthetic Hakka speech. The human portion includes approximately 70 hours of Dapu and Zhaoan recordings provided by the competition organizers. To expand this limited resource, we used YourTTS to generate around 310,000 utterances for each dialect, from which 25,000 per dialect were sampled to form a balanced synthetic training set. In addition, about 28,000 utterances targeting coarticulated syllables were synthesized to enhance phonological coverage, yielding roughly 100,000 synthetic training samples in total.

Metrics

We evaluated performance using two complementary metrics CER for character track and WER for pinyin track. These metrics jointly capture both graphemic and phonemic aspects of recognition, providing a comprehensive assessment of Hakka ASR performance.

Finetune Details

We fine-tuned the Whisper large-v2 model (Radford et al., 2022) for Hakka ASR using AdaLoRA. The adaptation targeted key modules including `k_proj`, `q_proj`, `v_proj`, `out_proj`, `fc1`, and `fc2`. Training was conducted for 25 epochs with the AdamW optimizer, a learning rate of 1×10^{-4}, and mixed-precision optimization (fp16).

To enhance robustness, several data augmentation strategies were applied with explicit probabilities and parameter ranges. Time stretching was applied with a probability of 0.25, adjusting the speaking rate within a range of 0.9 to 1.1 (Ko et al., 2015; McFee et al., 2015). Pitch shifting was also used with a probability of 0.25, modifying the fundamental frequency by up to ± 4 semitones (Ko et al., 2015; McFee et al., 2015). To simulate far-field effects, air absorption was introduced with a probability of 0.5, modeling distances between 10 and 50 meters (Habets, 2006). Environmental reverberation was added using impulse responses from RIR at a probability of 0.25 (Habets, 2006). Short environmental noises from the ESC-50 (Piczak, 2015) dataset were injected with a probability of 0.75, using signal-to-noise ratios (SNR) between 3 and 30 dB and durations of 2–8 seconds. Finally, Gaussian noise was added with a probability of 0.25, with SNR ranging from 5 to 40 dB (Ko et al., 2015).

Baselines

For the baseline system, we fine-tuned the Whisper large-v2 model directly on the original Hakka corpus provided by the organizers, which consists of Zhaoan and Dapu recordings. No synthetic speech data or augmentation techniques were applied. The model was trained under the same setup as our proposed method, using AdaLoRA for parameter-efficient fine-tuning with 25 training epochs, the AdamW optimizer, a learning rate of 1×10^{-4}, and mixed-precision optimization (fp16). This baseline serves as a reference point to evaluate the effectiveness of our data augmentation and synthetic data generation strategies.

4.2 Main Results

Table 1 summarizes the effectiveness of our proposed strategies. In low-resource and cross-dialect settings, ASR models often lack sufficient acoustic variability to generalize to real-world scenarios. To address this limitation, we deliberately designed and introduced a series of data augmentation techniques, including noise injection, utterance concatenation, time stretching, pitch shifting, air absorption, and environmental reverberation. These augmentations explicitly enriched the acoustic variability of the training corpus, thereby equipping the model with the ability to cope with noisy conditions, diverse speaking rates, phonological variation, and far-field speech scenarios. In addition, we incorporated TTS-generated synthetic data to compensate for limited phonological coverage and to expand the overall training corpus. As a result of combining enhanced acoustic diversity with more comprehensive phonological coverage, the model achieved substantial performance gains in cross-dialect recognition. Compared with the baseline trained solely on the original corpus (CER = 31.9%, WER = 51.5%), the final system reduced the character CER to **5.6%** and the pinyin WER to **16.7%**. These reductions—over fivefold in CER and more than threefold in WER—clearly demonstrate that synthetic data generation, together with carefully designed augmentation, provides the critical capabilities necessary for improving cross-dialect ASR robustness under low-resource conditions.

System	**CER (%)**	**WER (%)**
Baseline	31.9	51.5
Proposed Method	**5.6**	**16.7**

Table 1: Recognition performance on Hakka ASR under different training setups. The incorporation of TTS-generated synthetic data and augmentation methods yields substantial improvements over the baseline system

4.3 Error Analysis and Re-sampling

We performed detailed error analysis to identify persistent recognition errors. To address these, additional utterances were re-sampled from the 310,000 TTS-generated dataset for targeted fine-tuning. However, this error-driven re-sampling did not yield further performance gains, suggesting that improvements depend more critically on data quality and phonological coverage than on the sheer quantity of synthetic speech.

5 Conclusion

This work tackled the challenge of Hakka ASR under low-resource and dialectal variation by fine-tuning Whisper with LoRA and augmenting data with synthetic speech. Our approach reduced CER from 31.9% to 5.6% and WER from 51.5% to 16.7%, demonstrating that parameter-efficient adaptation with augmentation can yield substantial cross-dialect gains. Beyond Hakka, the framework offers a replicable path for extending large-scale ASR to other low-resource languages. Future work will target broader dialect coverage, more natural synthetic speech, and cross-lingual transfer, further supporting the preservation of endangered languages in the era of large-scale AI.

References

Nicholas J. Bryan. 2019. Impulse response data augmentation and simulation as alternatives to rir collections for speech recognition. In *Proceedings of WASPAA*, pages 229–233.

Edresson Casanova, Juliano Weber, Christopher Shulby, Antonio Junior, Rodolfo da Silva, Moacir Antonelli Ponti, Sandra Aluisio, Junichi Yamagishi, Yossi Adi, Mohamed Haidar, and 1 others. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning (ICML)*.

Emanuël AP Habets. 2006. Room impulse response generator. Technical report, Technische Universiteit Eindhoven. Technical Report.

Wei-Ning Hsu, Benjamin Bolte, Yung-Sung Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2020. Meta learning for low-resource speech recognition. In *ICASSP*.

Ye Jia, Heiga Zen, Ron J Weiss, and 1 others. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP*.

James M. Kates and Eugene J. Brandewie. 2020. Adding air absorption to simulated room acoustic models. *The Journal of the Acoustical Society of America*, 148(5):EL408–EL413.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *ICML*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Interspeech*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *Proceedings of ICASSP*, pages 5220–5224.

Bo Li, Abdelrahman Mohamed, and Geoffrey Zweig. 2020. Training data augmentation for end-to-end speech recognition using text-to-speech synthesis. In *ICASSP*.

X Liu, Jonas Pfeiffer, Sebastian Ruder, and 1 others. 2022. Adapterfusion: Non-destructive task composition for transfer learning. In *EACL*.

Brian McFee, Colin Raffel, Daniel Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018. ACM.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *NeurIPS*.

Andrew Rosenberg, Bhuvana Ramabhadran, Abhinav Sethy, and 1 others. 2019. Speech synthesis for data augmentation in noisy speech recognition. In *ICASSP*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, and 1 others. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*.

Jelle van der Meer. 2022. Evaluating the use of pitch shifting to improve automatic speech recognition. Master's thesis, Delft University of Technology.

Changhan Wang, Wei-Ning Hsu, and 1 others. 2021. Improving low-resource speech recognition with cross-lingual self-supervised learning. In *ICASSP*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*.

Muhammad Zahid and Imran Qazi. 2025. Pitch-speed feature space data augmentation for automatic speech recognition improvement in low-resource scenario. *International Journal of Speech Technology*.

Qianxi Zhang, Zhenheng Yang, Tianlong Chen, and Zhangyang Wang. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*.

應用 Whisper 與拼音後處理的客語語音辨識
Hakka Speech Recognition with Whisper and Pinyin Post-processing for FSR-2025

Chia-Hsin Lee*, **Yung-Jun Chang***, **Jin-Yan Wu***, **and Kuan-Yu Chen**
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
{d11415004, m11315045, m11315085, kychen}@mail.ntust.edu.tw

Abstract

本研究為參加 FSR-2025 客語語音辨識挑戰賽（Hakka ASR II）的技術報告，旨在推進客語自動語音辨識技術的發展。由於客語屬於低資源語言，且存在多種腔調，語音辨識面臨高度挑戰。我們以 Whisper-large-v2 為骨幹模型，設計兩階段訓練流程：首先利用「Hakka Across Taiwan (HAT)」語料庫進行模型調適，以捕捉客語的一般聲學特徵；其次在賽事方提供的 60 小時腔調語料上進行微調，以增強對目標資料的適應性。實驗發現，直接輸出客語漢字可達到良好的字錯率（CER），但由於腔調差異與拼音規則變化多，拼音任務表現顯著下降。為解決此問題，我們以漢字模型的編碼器初始化拼音模型，並提出結合 RoBERTa 漢字轉拼音、腔調判斷與字典修正的後處理模組，期望可以在比賽中提升辨識的成效。

1 Introduction

FSR-2025 客語語音辨識挑戰賽（Hakka ASR II）旨在推動台灣客語自動語音辨識技術，資料同時涵蓋朗讀與自然口語場景，並分成漢字與拼音兩種辨識任務，評估指標分別採用字元錯誤率（charactor error rate, CER）與音節錯誤率（syllable error rate, SER）。開發高品質客語語音辨識系統（automatic speech recognition, ASR）具備三方面意義：**(1)** 語言保存：完整蒐集與標註各腔調語料，有助文化傳承與方言研究；**(2)** 數位包容：提供客語族群友善的語音互動介面；**(3)** 研究價值：客語的多調系統與低資源特性，是檢驗新型 ASR 方法的重要試金石。

本研究以 Whisper-large-v2 為骨幹，我們首先比較使用不同的訓練資料量對於後續實驗成效的影響。在這部分的研究中，我們發現在客語漢字以及客語拼音任務中，單純提高訓練資料的規模會有不同結果。所以我們後續採用漢字模型的 Encoder 作為模型的初始權重，發現可以更穩定的提升客語拼音的成效。此外，雖然 Whisper 在進行客語漢字的輸出時能獲得不錯的表現，但由於客語屬於低資源語言，且存在多種腔調，這造成了 Whisper 在輸出客語拼音時準確率明顯下降。為了解決此一問題，我們提出一套後處理方法，期望可以在比賽中提升辨識的成效。

*These authors contributed equally to this work

2 Whisper

Whisper (Radford et al., 2023) 為 OpenAI 所提出，採用 Transformer Encoder-Decoder 架構的語音辨識模型。Whisper 基於 weakly supervised 的策略，以多達 680K 小時的語音資料進行訓練，這使得 Whisper 在多語言、多任務乃至於噪聲環境下的各種任務均具有強大的魯棒性（robustness）。得益於上述諸多的優點，許多開發低資源語音辨識模型的研究者也傾向使用 Whisper 作為預訓練模型。

利用 Whisper 進行單一語言的訓練最常見的策略是全參數微調，不過對於資源缺乏的語言，容易產生過擬合的問題。有鑑於此，近期對於資源缺乏的語言多採用 LoRA、Prompt tuning 的策略 (Qian et al., 2024)，僅訓練少量的參數，讓 whisper 可以學習新語言的特徵。除此之外，透過數據增強來增加訓練資料也是常見的方法，由於文字語料的取得難易度遠低於語音語料，以 TTS 等語音合成策略增加可用的訓練資源也行之有年 (Gokay and Yalcin, 2019)。我們的實驗將以 Whisper 預訓練模型作為骨幹架構，此外為了追求最佳辨識率，我們採用全參數微調的策略，以盡可能利用全部的參數。

3 方法

為了讓 Whisper 學習客語語音辨識任務，我們採用兩階段的訓練流程。在第一階段中，我們利用臺灣客語語音資料庫（HAT，Hakka Across Taiwan）(Liao et al., 2023) 對 Whisper

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 454–458
November 20-22, 2025 ©2025 Association for Computational Linguistics

#	Models	Dataset		CER (%)	
		pretraining	finetuning	dev	test
1	large-v2		賽事訓練語料	1.5	42.8
2	large-v2		HAT	1.3	35.1
3	large-v2	HAT	賽事訓練語料	1.0	19.8

Table 1: 客語漢字辨識結果（dev：錄製語料；test：媒體語料）

#	Models	Dataset		WER (%)	
		pretraining	finetuning	dev	test
1	large-v2		賽事訓練語料	6.9	53.0
2	large-v2		HAT	7.9	90.1
3	large-v2	HAT	賽事訓練語料	5.7	26.6

Table 2: 客語拼音辨識結果（dev：錄製語料；test：媒體語料）

模型進行初步訓練，以便模型能夠學習客語語音的基本特徵。接著，我們利用賽事方所提供約 60 小時的訓練語料進行微調，以提升模型對目標資料的適應能力。最後，我們對模型輸出的辨識結果進行後處理，並將其作爲最終的系統輸出提交。

3.1 語音辨識模型訓練

我們使用 Whisper large-v2 進行訓練，該模型爲各 32 層的 Transformer Encoder-Decoder 架構，約 1.5B 的參數量。前置訓練中我們以臺灣客語語音資料庫 (Hakka Across Taiwan，HAT) 語料庫，及賽事方提供共 60 小時之大埔與詔安腔作爲訓練資料。HAT 資料集是中華計算語言學學會推出之語料庫，其中包含豐富的語音標註資料，包含錄製語音、媒體語音以及語音合成的語音，腔調主要以海陸以及四縣腔爲主，此外我們使用 SpecAugmentation (Park et al., 2019) 進行資料增強。我們訓練兩個 Whisper 模型，分別輸出客語漢字與拼音兩種結果，以對應比賽的兩種任務。驗證集（development set）和測試集（test set）分別是熱身賽釋出的錄製語料及媒體語料。作爲額外的實驗，我們另行訓練了僅以大埔與詔安腔作爲訓練資料的模型進行比較。

我們的模型是使用 ESPNet toolkit (Watanabe et al., 2018) 進行實現，詳細的實驗設置列舉如下：

- Model: Whisper large-v2
- Learning rate: 5×10^{-5}
- Optimizer: AdamW
- Epoch: 50
- Save strategy: Top3

每個模型皆共訓練 50 epoch，保留成績最佳的三個 checkpoints 作爲最終模型。

3.2 結果比較

Tables 1和 2呈現我們的實驗結果。我們首先發現，在客語漢字以及拼音的場合中，單純提高訓練資料的規模會造成不同結果。客語漢字的 CER 隨著語料的增加有顯著的改善(見 Table 1，編號 1、2)；然而對於客語拼音而言，資料的增加反而造成 WER 的提升 (見 Table 2，編號 1、2)。我們推測此現象來自於客語拼音的規則複雜。除了對 Whisper 的 Decoder 而言，從頭學習客語拼音的難度高於客語漢字外，根據客家委員會推出之《客家語拼音方案》所描述，異於客語漢字，不同客語拼音對同一語句的拼音規則也因腔調、聲調而異，導致模型學習的困難。

綜合上面論述，爲了利於模型學習客語拼音，第二階段的模型訓練我們以漢字模型的 Encoder 作爲模型的初始權重，幫助 Decoder 能夠更穩定的學習客語拼音的規則。也就是說，Table 1中的編號 3 是由編號 2 模型繼續進行調適而成。Table 2中編號 3 的系統，是使用 Table1中編號 2 的 Encoder 與 Table 2中編號 2 系統的 Decoder 串接後再進行調適而得。

模型的第二階段，我們透過先前已學習客語語音基本特徵的 Whisper large-v2 模型進行再訓練，可見儘管初步訓練中使用的語料多爲海陸及四縣腔，仍有助於模型更好的掌握語音資訊，在以大埔與詔安腔爲主的驗證集與測試集中漢字與拼音均取得最優的結果 (見 Table 1、2，編號 3)。此外，客語拼音方面，成績的進步也證實了若 Encoder 能夠持續提供較優的語音特徵，將有助於 Decoder 學習更爲複雜的拼音規則。

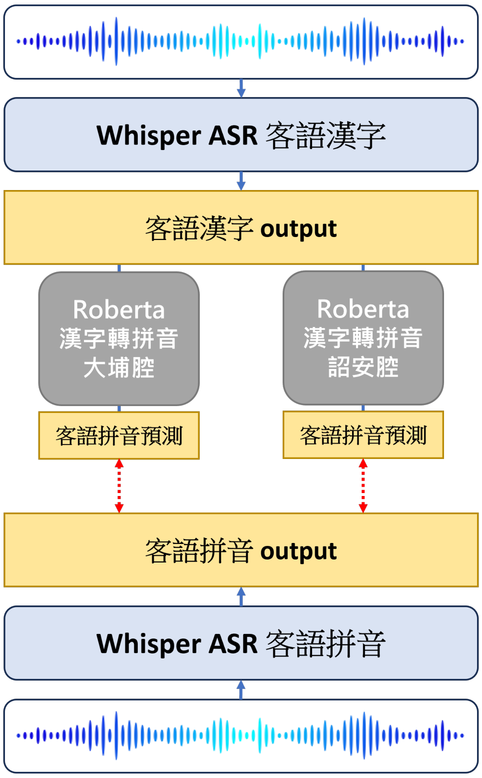


Figure 1: 整體流程：Whisper ASR 先輸出漢字，再透過 RoBERTa 漢字轉拼音模型輸出不同腔調的拼音，最後根據與 ASR 拼音的比較結果判斷腔調。

3.3 爲何需要後處理

客語屬於低資源語言，且存在多種腔調。在我們的實驗中發現，Whisper ASR 在輸出客語漢字時能獲得不錯的表現，但在直接輸出客語拼音時準確率明顯下降。原因在於同一個漢字，可能因腔調不同而對應到不同的拼音。因此，我們提出一套基於漢字的腔調判斷與拼音修正流程，以提升最終拼音輸出的準確度。系統流程如 Figure 1所示。

3.4 方法流程

3.4.1 漢字轉拼音模型訓練

我們首先依照腔調（如大埔腔、詔安腔）將資料集分開，並整理爲「漢字—拼音」配對資料。透過自訂的資料處理程式，將每個字對應到相應的拼音標籤，再使用 RoBERTa 預訓練模型 (Liu et al., 2019)，進行逐字的標記分類訓練，讓模型學會將漢字正確轉換爲拼音（含聲調）。每種腔調獨立訓練一個模型。

我們使用的中文預訓練 RoBERTa[*] 模型，是已經針對中文語料進行擴展式訓練 (Cui et al., 2021)。內容包含：

- **Masked Language Modeling**：隨機遮罩輸入中的部分字詞，並讓模型學習預測被遮蔽的字。
- **Whole Word Masking, (WWM)**：對於多字組成的詞彙（例如「臺灣科技大學」），遮罩時會同時遮罩整個詞，而非僅遮罩單一字，提升模型學習詞級語義的能力。

因此，在中文任務上（如詞性標註、序列標記、問答任務等）可以展現出比傳統 BERT 更優異的表現。

基於這個中文預訓練 RoBERTa 模型，我們採用 HuggingFace Transformers 框架進行訓練，以成爲漢字轉拼音模型。具體設置如下：

- Model：`hfl/chinese-roberta-wwm-ext`
- Learning rate：5×10^{-5}
- Optimizer：AdamW（使用 HuggingFace 預設）

3.4.2 腔調判斷

至此，我們已具備三種模型：

1. Whisper ASR 直接輸出拼音。
2. Whisper ASR 輸出漢字。

[*] https://huggingface.co/hfl/chinese-roberta-wwm-ext

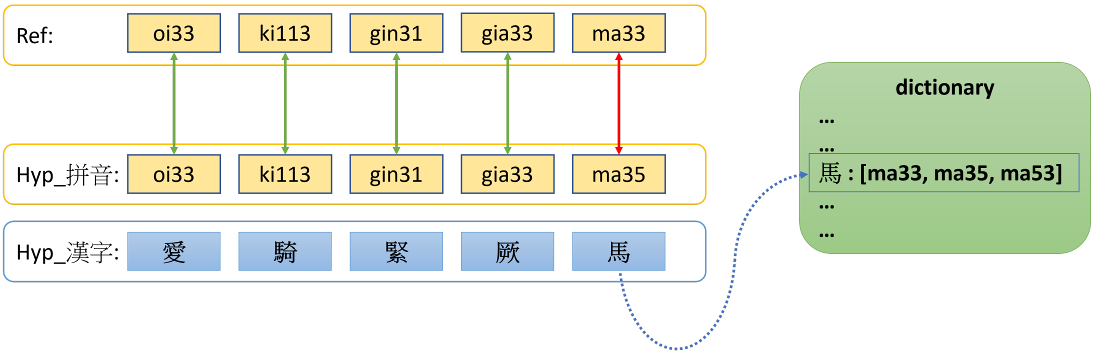


Figure 2: 字典修正流程：alignment 偵測到替換錯誤（例：「ma35」對「ma33」），再由字典提供候選拼音，選擇與參考輸出最相近者作修正。

Table 3: 熱身賽資料集媒體語料的 WER 結果比較。

系統	WER
Whisper ASR（直接拼音）	0.265
Whisper ASR（漢字 → 拼音）	0.266
＋ 字典修正（最終）	**0.264**

3. 各腔調的漢字轉拼音模型。

我們首先採用 Whisper ASR 辨識輸出漢字，再分別利用大埔腔與詔安腔的漢字轉拼音模型轉換成拼音序列，並與 Whisper ASR 辨識輸出的拼音進行 WER 計算。我們選擇 WER 較低的拼音，做爲這個語句的腔調。

3.4.3 字典比對修正

即便在同一腔調中，單一漢字仍可能對應到多種拼音，導致模型選擇錯誤。因此，我們整理出一份字典，將每個漢字對應到其在資料中出現過的所有拼音變體。在後處理過程中，若對齊 (alignment) 結果中出現替換錯誤 (Substitution Error)，我們會從字典中取出候選拼音，逐一比較 CER（字元錯誤率），並選擇最接近參考輸出的拼音進行修正。Figure 2展示了字典修正流程。

3.5 實驗結果

3.5.1 資料集

我們使用熱身賽資料中錯誤率較高的媒體語料進行實驗，並以 WER（詞錯誤率）作爲評估指標。

3.5.2 結果比較

實驗結果如 Table 3 所示。Whisper 直接輸出拼音的 WER 爲 0.265。透過漢字轉拼音並進行腔調判斷後，WER 爲 0.266。最後加入字典修正機制後，WER 降至 0.264。

4 Acknowledgment

This work was supported by the National Science and Technology Council (NSTC) of Taiwan under Grants NSTC 112-2636-E-011-002, NSTC 112-2628-E-011-008-MY3, and NSTC 113-2640-B-002-005. Additional support was provided by the "Empower Vocational Education Research Center" at the National Taiwan University of Science and Technology (NTUST) through the Featured Areas Research Center Program, as part of the Higher Education Sprout Project funded by the Ministry of Education (MOE), Taiwan. The authors also thank the National Center for High-Performance Computing, National Applied Research Laboratories (NARLabs), Taiwan, for providing essential computational and storage resources.

References

Yiming Cui, Wanxiang Che, and Ting Liu et al. 2021. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:2103.00492*.

Ramazan Gokay and Hulya Yalcin. 2019. Improving low resource turkish speech recognition with data augmentation and tts. In *2019 16th International Multi-Conference on Systems, Signals Devices (SSD)*, pages 357–360.

Yuan-Fu Liao, Shaw-Hwa Hwang, You-Shuo Chen, Han-Chun Lai, Yao-Hsing Chung, Li-Te Shen, Yen-Chun Huang, Chi-Jung Huang, Hsu Wen Han, Li-Wei Chen, Pei-Chung Su, and Chao-Shih Huang. 2023. Taiwanese hakka across taiwan corpus and formosa speech recognition challenge 2023 - hakka asr. 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.

Yinhan Liu, Myle Ott, Naman Goyal, and Jingfei Du et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, pages 2613–2617.

Mengjie Qian, Siyuan Tang, Rao Ma, Kate M. Knill, and Mark J.F. Gales. 2024. Learn and Don't Forget: Adding a New Language to ASR Foundation Models. In *Interspeech 2024*, pages 2544–2548.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. In *Interspeech 2018*, pages 2207–2211.

A Study on a Low-Resource Speech Recognition System for Taiwan Hakka Based on Whisper and LoRA
(基於 Whisper 與 LoRA 的低資源模型之臺灣客家語語音辨識系統研究)

劉政廷
National Yunlin University of Science and Technology Department of IM (Program in Artificial Intelligence) / 雲林縣斗六市大學路三段 123 號
zhengtingliu0104@gmail.com

王恆佑
National Yunlin University of Science and Technology Department of IM (Program in Artificial Intelligence) / 雲林縣斗六市大學路三段 123 號
yoyo1234870@gmail.com

廖宜翔
National Yunlin University of Science and Technology Department of IM (Program in Artificial Intelligence) / 雲林縣斗六市大學路三段 123 號
sean0908699966@gmail.com

邱重袁
National Yunlin University of Science and Technology Department of IM (Program in Artificial Intelligence) / 雲林縣斗六市大學路三段 123 號
alan1109135@gmail.com

黃昭義
National Yunlin University of Science and Technology Department of IM (Program in Artificial Intelligence) / 雲林縣斗六市大學路三段 123 號
joyh@yuntech.edu.tw

摘要

本研究旨在開發一套高效能的臺灣客語自動語音辨識（ASR）系統，以應對客語作為低資源語言所面臨的保存與數位化挑戰。本研究採用 whisper-large-v3-taiwanese-hakka 作為基礎模型，此模型基於先進的 Transformer 編碼器-解碼器架構。為達成參數高效且能適應新語言的目標，我們採用了 LoRA (Low-Rank Adaptation) 微調策略，並特別針對模型中 q_proj、k_proj、v_proj、out_proj、fc1、fc2 等多個關鍵模組進行適配。實驗結果表明，相較於基礎模型在 FSR-2025 HAT-Vol2 測試集上 23.07% 的字元錯誤率 (CER)，經過 LoRA 微調後的模型表現出色，最終將 CER 顯著降低至 7.07%。訓練過程監控顯示，模型的驗證集損失與錯誤率皆穩定下降並收斂，證明 LoRA 策略能在不發生毀滅性遺忘的前提下，成功地將大型模型的知識遷移至客語辨識任務，提供了一個高效的解決方案。

Abstract

This study aims to develop a high-performance Automatic Speech Recognition (ASR) system for Taiwan Hakka, addressing the preservation and digitalization challenges it faces as a low-resource language. We utilized whisper-large-v3-taiwanese-hakka as the base model, which is built upon an advanced Transformer encoder-decoder architecture. To achieve parameter-efficient adaptation for the new language, we employed the Low-Rank Adaptation (LoRA) fine-tuning strategy, specifically adapting key modules within the model, including q_proj, k_proj, v_proj, out_proj, fc1, and fc2. The experimental results demonstrate outstanding performance. Compared to the base model's Character Error Rate (CER) of 23.07% on the FSR-2025 HAT-Vol2 test set, the LoRA-tuned model achieved a significant reduction, bringing the final CER down to 7.07%. Monitoring of the training process showed that the model's validation loss and error rate both steadily decreased and converged. This confirms that the LoRA strategy can successfully transfer knowledge from the large model to the Hakka recognition task without suffering from catastrophic forgetting, providing an efficient solution.

關鍵字：客家語、自動語音辨識
Keywords: Hakka, Automatic Speech Recognition

1 前言

自動語音辨識（Automatic Speech Recognition, ASR）技術在高資源語言（如英語、漢語普通話）的應用已臻成熟，並廣泛

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 459–466
November 20-22, 2025 ©2025 Association for Computational Linguistics

部署於智慧助理、語音輸入及智慧家居等場域。近年來，此領域的突破主要歸功於深度學習，特別是 Vaswani et al. (2017) 提出的 Transformer 架構，其顯著提升了語音辨識的準確率與泛化能力。然而，對於全球眾多資源匱乏的低資源語言而言，ASR 系統的發展因語料不足而依然面臨嚴峻挑戰。

臺灣客語作為重要的本土語言之一，便面臨著語料稀缺且使用人口逐年下降的雙重困境。為此，發展高效能的客語 ASR 系統，不僅是技術層面的探索，更對語言的數位保存、教育推廣與文化傳承具有深遠的學術與社會意義。

在此背景下，「2025 Formosa 語音辨識挑戰賽」提供了標準化的 HAT-Vol2 客語語料庫，為本研究奠定了關鍵的實驗基礎。同時，OpenAI 的 Whisper 模型（Radford et al., 2023）透過在海量多語言資料上的預訓練，展現了在低資源情境下卓越的遷移潛力。因此，本研究的核心動機便是結合前述契機，將先進的大型預訓練語音模型應用於客語辨識任務，並透過微調策略探索其效能極限，以期為客語數位化工程提供一個穩固的技術方案。

2 文獻回顧

為奠定本研究之理論基礎與確立研究定位，本章將回顧自動語音辨識技術的發展脈絡、探討其在低資源語言上面臨的挑戰，並檢視臺灣本土語言相關研究之現況。首先，我們將追溯 ASR 技術的演進，從早期的統計模型（如隱馬可夫模型(Hidden Markov Models, HMMs)）發展至現今由深度學習主導的端到端架構，特別是 Transformer 模型的崛起如何革新了整個領域。

2.1 自動語音辨識技術的演進

ASR 的發展可追溯至 20 世紀 50 年代早期，當時的系統主要依賴基於動態時間規劃 (Dynamic Time Warping, DTW) 的方法來進行模式比對。隨著統計學與機器學習的進步，HMMs 成為主流，並在 1980 至 1990 年代主導了語音辨識的研究方向。HMMs 能夠有效地建模語音信號的時間序列特性，並與高斯混合模型 (Gaussian Mixture Models, GMMs) 結合，形成了經典的 GMM-HMM 框架，廣泛應用於語音辨識系統中。

進入 2010 年後，深度學習技術的興起帶來了革命性的改變。深度神經網路 (Deep Neural Networks, DNNs) 被引入到聲學建模中，逐漸取代傳統的 GMM。接著，卷積神經網路 (Convolutional Neural Networks, CNNs) 與長短期記憶網路 (Long Short-Term Memory, LSTM) 更進一步提升了模型的表現，使得 ASR 系統在大型語料下取得顯著突破。2017 年，Vaswani 等人提出的 Transformer 架構，首次完全拋棄循環結構，透過自注意力機制 (Self-Attention Mechanism) 同時捕捉長距依賴與上下文資訊，顯著提升了序列建模能力。此後，基於 Transformer 的模型（如 Conformer、wav2vec 2.0）逐漸成為主流，並奠定了現今端到端語音辨識系統的基礎。

2.2 低資源語言的 ASR 挑戰

儘管深度學習推動了語音辨識的快速進展，但這些成功主要集中於高資源語言，如英語與漢語普通話。低資源語言面臨兩大挑戰：

1. 語料不足：缺乏大規模且標註完善的語音－文字對齊資料，使得模型難以進行有效訓練。
2. 語言特性複雜：許多低資源語言存在方言差異、口語化強烈或文字系統尚未標準化等問題。

為克服這些困境，研究者提出了多種方法：

- 跨語言遷移學習 (Cross-lingual Transfer Learning)：先在高資源語言進行預訓練，再將模型微調於低資源語言。
- 多語言訓練 (Multilingual Training)：同時使用多種語言進行訓練，以共享跨語言特徵。

- 參數高效微調 (PEFT)：如 LoRA (Hu et al., 2022)，僅調整部分權重即可快速適應新語言，降低計算成本與記憶體需求。

這些方法已在多種低資源語言上展現成效，例如 Meta 的 BABEL 計畫（Harper, 2014）針對 26 種語言建立了語音資源，wav2vec 2.0 也在非洲語言與南亞語言上獲得成功應用。

2.3 臺灣本土語言的語音研究

臺灣的語言多樣性極為豐富，包括閩南語、客家語、原住民族語言等。然而，相關的語音辨識研究仍處於起步階段。過去研究多集中於閩南語，例如透過 Kaldi 工具建立 GMM-HMM 與 DNN-HMM 系統；而客語因語料稀缺與方言差異，研究進展較為有限。近年來，隨著開放資料集（如 FSR 挑戰賽 HAT-Vol2）的釋出，客語 ASR 的研究基礎逐漸建立，為本研究提供了重要契機。

3 研究方法

為開發高效能的臺灣客家語語音辨識系統，本研究採用了基於大型預訓練模型的監督式微調技術路徑。本章將詳細闡述整體實驗設計與流程。首先，我們將介紹核心採用的 Whisper-large-v3 模型之架構基礎，說明其為何適合作為低資源語言辨識的起點。接著，將闡述本研究採用的參數高效微調（Parameter-Efficient Fine-Tuning, PEFT）策略，特別是 LoRA 技術的原理與具體設定，此策略旨在以最少的計算資源達成最佳的模型適應性。隨後，將說明針對 FSR-2025 所提供的 HAT-Vol2 語料庫所進行的資料前處理步驟，以確保資料品質與模型訓練的穩定性。最後，將定義本研究用以衡量模型性能的評估指標，包括字元錯誤率（CER）與詞錯誤率（WER）。

3.1 模型微調策略

在低資源語言的設定下，對大型模型進行完整的參數再訓練（Full Fine-tuning）成本極高。因此，本研究選擇了 PEFT 中的 LoRA 策略。LoRA 的核心思想是在原有的預訓練權重旁，額外引入少量可訓練的低秩矩陣來模擬權重的更新，如此便能在保留模型預訓練知識、避免災難性遺忘的同時，大幅降低計算需求與記憶體佔用。

為了讓模型能更全面地適應客語的聲學與語言特性，我們參考了相關研究 並擴展了 LoRA 的目標模組。根據我們的訓練腳本，本研究特別針對 Transformer 架構中的六個關鍵模組進行微調，其 target_modules 參數設定如下：q_proj, k_proj, v_proj, out_proj, fc1, fc2。

選擇這些模組的原因在於，它們是 Transformer 模型的核心組成部分：

- q_proj、k_proj、v_proj、out_proj 是多頭自注意力機制中的關鍵線性投射層，微調這些層有助於模型學習如何更好地關注與客語聲學特徵相關的資訊。
- fc1 與 fc2 則是前饋神經網路中的兩個線性層，負責對注意力機制提取的特徵進行非線性轉換與更高層次的表徵學習。

透過對這六個模組進行適配，模型能夠在注意力層面與特徵轉換層面同時進行調整，從而更有效地將其預訓練知識遷移並適應於客語的獨特結構。

3.2 資料前處理細節

客語 ASR 面臨的挑戰之一在於語音－文字對齊與方言差異，因此我們進行了以下處理：

- 音訊正規化：統一採樣率至 16kHz，移除背景噪音過重的片段。
- 斷詞處理：由於客語標註多以「字」為單位，我們採取字元級 (character-level) 輸出，以降低斷詞不一致的影響。
- 混語現象處理：客語常混雜華語或英語詞彙，我們保留此特性，以反映真實語境。

3.3 評估指標 (Evaluation Metrics)

本研究採用業界標準的詞錯誤率 (WER) 與字元錯誤率 (CER) 作為評估模型性能的指標。計算公式為$\frac{S+D+I}{N}$。其中 S 為替換 (Substitutions) 錯誤數，D 為刪除 (Deletions) 錯誤數，I 為插入 (Insertions) 錯誤數，N 為參考文本的總詞數（用於 WER）或總字數（用於 CER）。

4 結果與討論

雖然微調後的 whisper-large-v3-taiwanese-hakka 在臺灣客語 ASR 任務上展現了良好性能（7.07% CER 與 40.99% WER），但進一步的錯誤分析揭示了模型在語音辨識上的挑戰。

4.1 實驗結果

本研究將採用 LoRA 微調後的 whisper-large-v3 模型應用於臺灣客家語自動語音辨識任務。相較於未使用 LoRA 微調前的基準表現（CER 23.07%, WER 78.15%），經過 LoRA 微調後的模型在 FSR-2025 HAT-Vol2 測試集上取得了顯著的進步，最終達到了 7.07% 的字元錯誤率與 40.99% 的詞錯誤率。此結果驗證了大型多語言預訓練模型結合微調策略，能在低資源語言上達到高效能。

4.2 訓練過程分析

模型的訓練過程展現出良好的收斂性。如圖 1 所示，訓練損失 (loss) 與驗證損失 (eval_loss) 均隨全局步驟 (global_step) 穩定下降。同樣地，如圖 2 所示，驗證集詞錯誤率 (eval_wer) 亦呈現明顯的下降趨勢，並在訓練後期趨於平穩。我們觀察到，由於 Whisper-large-v3 模型本身強大的預訓練能力，其在客語資料上的適應速度非常快，在訓練的早期階段（約 1000 步內）損失和錯誤率已大幅降低。因此，本研究採用了提前中止訓練的策略（設定最大步數為 4000 步），在不到一個完整 epoch 的訓練量下即完成微調。此舉不僅顯著節約了計算資源，亦有效降低了模型在相對有限的客語資料上產生過擬合的風險。

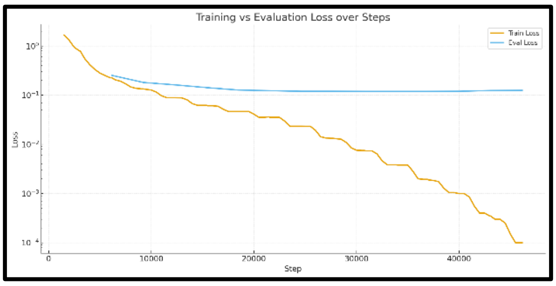


圖 1 CER 模型訓練與驗證損失曲線

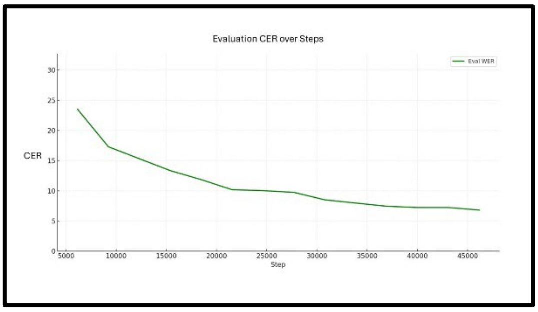


圖 2 驗證集詞錯誤率 (CER) 變化

4.3 錯誤分析與討論

雖然微調後的模型在臺灣客家語 ASR 任務上展現了良好性能，但進一步的錯誤分析揭示了模型在語音辨識上的具體挑戰。依據標準計算方式，我們將錯誤分為替換、刪除與插入三大類，其在測試集中的比例分佈如表 1 所示。

表 1：測試集中三種錯誤類型的比例表

錯誤類型	百分比 (%)	典型例子

替換 (S)	62.4	「路上」→「無上」
刪除 (D)	25.7	省略動詞，如「做」被忽略
插入 (I)	11.9	額外插入詞，如多餘的「啊」

顯示了三類錯誤的分布比例，可見替換錯誤佔據主要部分，說明聲學相似詞的辨識仍是系統的主要瓶頸。

4.4 替換錯誤

替換錯誤多數發生於聲學相似字，例如：

- 例一：參考句「你 頭 擺 知 路上 正 經 講 啊」，模型輸出「你 頭 擺 知 無上 正經 講 啊」。
- 例二：參考句「這 埕壁 真 滑 哦」，輸出為「這 埕壁 真 好 哦」。

這些錯誤顯示，客語同音或近音字的分辨度不足，尤其在缺乏上下文輔助時更為明顯。

4.5 刪除錯誤

刪除錯誤主要發生於長句，模型在遇到多重修飾詞或動詞堆疊時，傾向省略部分字詞。例如：

- 例一：參考句「做 得 恁 晊 恁 讚 會 焳 手頭 恁 順」，輸出為「得 恁 會 手頭 恁 順」。

這反映了模型在處理客語複合句與口語化表達時的挑戰。

4.6 插入錯誤

插入錯誤比例雖較低，但常見於句尾，模型傾向額外生成語氣助詞「啊」、「呢」，例如：

- 例一：參考句「今晡日 天氣 真 好」，輸出「今晡日 天氣 真 好 啊」。

此類錯誤與客語語音的語尾拖音、口語韻律有關。

4.7 成功與失敗案例

- 成功案例：短句如「五月節愛食粽」與「寒著愛戴嘴揞」，CER 達到 0%，顯示模型能準確捕捉簡短口語。
- 失敗案例：在長句「這埕壁真滑哦即算係對偓和偓輪嘛係莫擘心偲人做會到个斷雞作埒即會用啊」中，模型將「滑」→「好」、「輪」→「崙」、「雞」→「基」，出現多重替換錯誤。

4.8 實驗結果

本研究將 Whisper-large-v3-taiwanese-hakka 模型透過監督式微調應用於臺灣客家語自動語音辨識任務。在 FSR-2025 HAT-Vol2 測試集上，模型在未用 LoRA 微調前達到 23.07%的字元錯誤率與 78.15%的詞錯誤率，但是模型經過 LoRA 後最終達到 7.07% 的字元錯誤率與 40.99% 的詞錯誤率。並在訓練過程中驗證集損失與錯誤率皆穩定下降並成功收斂，顯示模型具備良好的收斂性與適應性。這些結果驗證了大型多語言預訓練模型結合微調策略，能在低資源語言上達到高效能。

4.9 結果分析

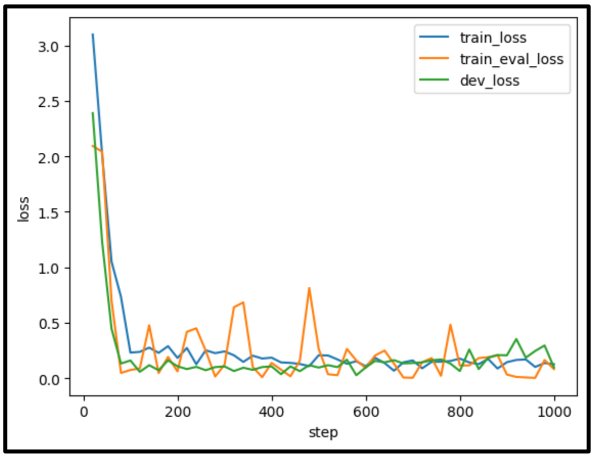


圖 3 CER 訓練與驗證損失曲線

```
[2025-09-14 12:06:21] [TRAIN] step  5820  loss=0.0396  lr=0.000049
[2025-09-14 12:13:19] [TRAIN] step  5840  loss=0.0719  lr=0.000049
[2025-09-14 12:20:16] [TRAIN] step  5860  loss=0.0776  lr=0.000048
[2025-09-14 12:27:14] [TRAIN] step  5880  loss=0.0295  lr=0.000048
[2025-09-14 12:34:11] [TRAIN] step  5900  loss=0.0492  lr=0.000048
[2025-09-14 12:41:13] [TRAIN] step  5920  loss=0.0773  lr=0.000047
[2025-09-14 12:48:11] [TRAIN] step  5940  loss=0.0503  lr=0.000047
[2025-09-14 12:55:10] [TRAIN] step  5960  loss=0.0377  lr=0.000047
[2025-09-14 13:02:08] [TRAIN] step  5980  loss=0.0456  lr=0.000046
[2025-09-14 13:09:08] [TRAIN] step  6000  loss=0.0463  lr=0.000046
```

圖 4 CER 訓練步驟

本研究透過訓練日誌對模型的學習過程進行了詳細分析。在本次實驗中，我們的訓練腳本設定了 --max_steps=6000（如圖 4）作為訓練中止的條件，此參數的優先級高於 epochs 設定，成為訓練長度的主要控制器。訓練日誌明確顯示，訓練最終停止於第 6000 步，此刻對應的 epoch 為 0.4。採用此策略的原因在於，我們發現 Whisper-large-v3-taiwanese-hakka 模型憑藉其強大的預訓練能力，在客語資料上的收斂速度極快。如圖 3 所示，訓練損失與驗證集錯誤率在訓練初期便已大幅下降並趨於穩定。因此，透過設定最大步數來精確控制訓練長度，使我們能在模型已達良好收斂狀態時及時中止訓練。此舉不僅有效地節約了大量的計算資源，也降低了模型在相對有限的客語資料集上產生過擬合的風險。

從最終輸出字串的錯誤類型進行分析，其比例分布為：替換佔 62.4%、刪除佔 25.7%、插入佔 11.9%。其中替換錯誤主要來自聲學相似字，顯示客語同音或近音詞的辨識是主要挑戰；刪除錯誤多出現在長句或多重修飾結構，反映模型在處理口語化語法時的不足；插入錯誤則常出現在句尾，多為語氣助詞，與客語韻律特徵相關。這些結果說明模型在短句表現穩定，但在長句則容易出現累積錯誤，且在訓練時資料有點偏向短句形式，很有可能模型會因為這樣學不到長句的特徵，甚至出現缺字的問題。

本研究突顯 Whisper-large-v3 在低資源語言上的遷移潛力，也證實參數高效微調策略能在有限語料下達到良好效果。然而，本研究仍受到語料規模與方言差異限制，聲學相似詞的辨識問題仍待解決。

5 結論與建議

本研究成功地將 whisper-large-v3-taiwanese-hakka 模型透過監督式微調應用於臺灣客家語語音辨識任務。

5.1 結論

本研究旨在處理臺灣客語的語音辨識問題，並透過對 whisper-large-v3-taiwanese-hakka 模型進行監督式微調，已於 CER 上取得顯著改進，成功將其由微調前的 23.07% 大幅降低至 7.07%。

儘管詞錯誤率仍有 40.99%，但錯誤分析顯示，其主要瓶頸源於客語中大量聲學特徵相似的同音或近音詞所導致的「替換錯誤」。此詞彙層級的挑戰，可俟未來客語詞彙庫與語言模型建置更臻完整後，再引入傳統統計模型（如 HMMs）或更先進的解碼技術逐步完善之。

從訓練過程來看，模型的損失函數與驗證集錯誤率均呈現穩定下降並成功收斂的趨勢，證明了微調策略的有效性。綜上所述，本研究成功驗證了將大型多語言模型遷移至特定低資源語言的可行性與巨大潛力，並為臺灣客語的語音技術發展提供了一個在字元辨識層級上高效且可靠的技術方案。

5.2 建議

基於本研究的發現與限制，我們提出以下幾點作為未來研究的建議方向：

1. 擴充與多樣化訓練語料： 本研究的性能仍受限於 HAT-Vol2 語料庫的規模。未來的研究應致力於收集更多元的客語語音資料，特別是增加長句與複雜句法的比例，以改善模型在處理長句時容易出現刪除錯誤的問題。同時，納入更多不同地區的腔調與口

音，有助於提升模型的泛化能力與穩健性。

2. 引入外部語言模型 (External Language Model)：目前模型的辨識主要依賴其內部學習到的語言知識。為了進一步降低因聲學混淆造成的替換錯誤（WER 偏高的主因），建議在解碼階段整合一個專門在大量客語文本上訓練的外部語言模型。透過 Beam Search with LM Fusion 等技術，語言模型可以提供更準確的詞彙機率，幫助模型在聲學上模糊不清時，做出更合理的詞彙選擇（例如，判斷「路上」比「無上」更為合理）。

3. 探索更先進的 PEFT 策略：本研究已證明微調的有效性。未來可進一步比較不同 PEFT 方法的效益，例如探索 LoRA 的不同配置（如調整 target_modules 或秩 r 的大小），或是導入 QLoRA 等更節省記憶體的技術，以在有限的計算資源下尋求最佳的性能與效率平衡點。

進行更細緻的錯誤歸因分析：建議未來可進行更深入的錯誤分析，例如針對不同腔調、語速或信噪比的語句進行分類錯誤統計。若能進行音素級別 (Phoneme-level) 的錯誤分析，將能更精準地定位模型在聲學上混淆的具體音素對，從而為模型結構的改進或資料增強策略提供更明確的指導方向。

References

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460. https://arxiv.org/abs/2006.11477

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. Advances in Neural Information Processing Systems, 32, 7059–7069. https://arxiv.org/abs/1901.07291

FSR Challenge. (2025). Formosa Speech Recognition Challenge 2025 (HAT-Vol2 Dataset). Retrieved from https://fsr2025.org

Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ... & Wu, Y. (2020). Conformer: Convolution-augmented transformer for speech recognition. Proceedings of Interspeech 2020, 5036–5040. https://doi.org/10.21437/Interspeech.2020-3015

Harper, M. (2014). Learning from 26 languages: Program management and science in the BABEL program. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 1–15. https://aclanthology.org/C14-1001

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. Proceedings of the 36th International Conference on Machine Learning (ICML 2019), 2790–2799. https://arxiv.org/abs/1902.00751

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Wang, S., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. arXiv Preprint. https://arxiv.org/abs/2106.09685

Hsieh, S. C., Huang, C. R., & Chen, K. J. (2013). Language resources for Taiwanese languages: Challenges and opportunities. Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC), 185–194.

Huang, C. R., & Hsieh, S. C. (2016). Corpus-based approaches to minority languages in Taiwan: Hoklo and Hakka. Language Resources and Evaluation, 50(3), 623–644. https://doi.org/10.1007/s10579-016-9352-2

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., Kahn, J., Likhomanenko, T., ... & Collobert, R. (2020). Massively multilingual ASR with large-scale weakly supervised data. Proceedings of Interspeech 2020, 4751–4755. https://doi.org/10.21437/Interspeech.2020-2837

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. arXiv Preprint. https://arxiv.org/abs/2212.04356

Tseng, H. H., Liu, C. L., Gao, Z. M., & Chen, K. J. (2002). A hybrid approach for automatic classification of Chinese unknown verbs [以構詞律與相似法為本的中文動詞自動分類研究]. International Journal of Computational Linguistics & Chinese Language Processing, 7(1), 1 – 28. https://aclanthology.org/O02-1001

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://arxiv.org/abs/1706.03762

Zhang, Y., Chen, W., Li, M., Wu, Y., & Liu, S. (2023). Benchmarking end-to-end speech recognition models for low-resource languages. arXiv Preprint. https://arxiv.org/abs/2305.10713

Zoph, B., & Knight, K. (2016). Multi-source neural translation. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 30–34. https://doi.org/10.18653/v1/N16-1004

A Compact Whisper+LoRA Baseline for Taiwanese Hakka ASR in FSR-2025

Hung-Ting Hsieh
Independent Researcher
hsiehk0214@gmail.com

Abstract

We present a compact baseline for the Formosa Speech Recognition (FSR-2025) Taiwanese Hakka ASR challenge. Our system fine-tunes *Whisper-large-v2* (Track 1) and *Whisper-large-v3-turbo* (Track 2) (Radford et al., 2022) with LoRA (Hu et al., 2021), using consistent text normalization and balanced dev splits, without external data or language models. On the official warm-up set, we obtain 10.94% CER for Track 1 (Hanzi) and 28.48% SER for Track 2 (Pinyin). We provide simple yet reproducible pipelines covering data preparation, training, inference, and evaluation. Code is available at github.com/Kevindic0214/FSR-Challenge-2025.

Keywords: Automatic speech recognition; Hakka; Whisper; LoRA; low-resource; FSR-2025; CER; SER

1 Introduction

Taiwanese Hakka is a low-resource language variant of significant cultural value. FSR-2025 defines two tracks: Track 1 evaluates character error rate (CER) on Hanzi, and Track 2 evaluates syllable error rate (SER) on Pinyin. We aim to provide a strong, minimal-requirement baseline using *Whisper-large-v2* (Track 1) and *Whisper-large-v3-turbo* (Track 2) fine-tuned with low-rank adaptation (LoRA), emphasizing practical engineering choices and reproducibility over model complexity.
In this work, we follow the official specification of the FSR-2025 challenge (FSR2025).

Contributions. (i) A reproducible baseline for both tracks with unified normalization that matches evaluation; (ii) a simple LoRA recipe runnable on a single 24 GB GPU with balanced speaker-based dev split; (iii) competitive warm-up results with lightweight error and length-bucket analyses.

Table 1: Dataset overview and evaluation split.

Split	Size	Notes
HAT-Vol2 (train)	∼60 h	Dapu/Zhao'an; 16 kHz mono
Warm-up (eval)	∼10 h / 4,299 utt	Official FSR-2025 set
Dev speakers	12 (balanced)	DF/DM/ZF/ZM allocation

2 Task and Data

We train on the HAT-Vol2 corpus (~60 hours; Dapu and Zhao'an dialects; 16 kHz mono) and evaluate on the FSR-2025 warm-up set (~10 hours; 4,299 utterances total). We build manifests via dedicated scripts for each track, apply Unicode NFKC normalization, remove zero-width characters, and adopt track-specific text processing: Hanzi cleaning for Track 1 and Pinyin digit-tone policy for Track 2. Dev speakers are selected in a balanced way across DF/DM/ZF/ZM groups for stable validation. We rely on the HAT-Vol2 dataset (HAT-Vol2) and the official warm-up set (FSR2025) for training and evaluation.

Normalization policy. Track 1 (Hanzi): apply NFKC, remove zero-width characters, map mixed punctuation to Chinese forms, and strip spaces to align with evaluation. Track 2 (Pinyin): apply NFKC, remove zero-width characters, map ü/ǘ/... and "u:/U:" to "v", keep only [a–z0–9] and single spaces, and by default drop starred syllables (e.g., "*ki53" or "ki53*"); an optional fix merges split-tone forms (e.g., "ki 53" → "ki53").

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 467–470
November 20-22, 2025 ©2025 Association for Computational Linguistics

3 Related Work

Low-resource ASR has been explored in multilingual programs such as Babel (Harper, 2014). Whisper (Radford et al., 2022) is a strong multilingual recognizer; we adapt it to Hakka via parameter-efficient fine-tuning. LoRA (Hu et al., 2021) reduces trainable parameters for seq2seq models while retaining quality, enabling practical fine-tuning on 24 GB GPUs.

4 Approach

We fine-tune *Whisper-large-v2* (Radford et al., 2022) with LoRA (Hu et al., 2021) (rank 16, α=32, dropout 0.05). Training uses gradient checkpointing, bf16 when available, and label smoothing. For Track 1 decoding, we force Chinese transcription via the decoder prompt; Track 2 uses language-appropriate decoding without language forcing. Beam search with 5 beams and temperature 0.0 is used unless specified.
Implementation details: we apply LoRA adapters to attention and MLP modules (`q_proj`, `k_proj`, `v_proj`, `out_proj`, `fc1`, `fc2`); enable TF32 for faster, stable training on recent GPUs; and use label smoothing of 0.1.

We keep Whisper's default suppression behavior (do not forcibly clear `suppress_tokens`), disable the generation cache during training, and enable early stopping on the dev metric (patience 2). bf16 is automatically used when supported; otherwise fp16 on GPU.

Implementation. We implement training and inference with HuggingFace Transformers and PEFT on PyTorch. Audio I/O uses `torchaudio` for Track 1 and `soundfile`+`librosa` for Track 2. Manifests are JSONL with fields {`utt_id`, `audio`, `text`/`hanzi`/`pinyin`, `group`}; relative audio paths are resolved via a root flag. During training we enable gradient checkpointing (non-reentrant when available), set `use_cache`=False, and turn on TF32. Decoding uses `num_beams`=5, `temperature`=0.0, `no_repeat_ngram_size`=3, `length_penalty`=1.0, and `max_new_tokens`=256; we force a Chinese decoder prompt only for Track 1. We log CER/SER, exact match, group/length-bucket scores, 3-gram repetition rate, throughput, and peak memory; training uses AdamW with a linear schedule and 500 warmup steps, and we save only the LoRA adapter and processor for lightweight deployment.

Track	Metric	Score
Track 1 (Hanzi)	CER / EM	10.94% / 58.06%
Track 2 (Pinyin)	SER / EM	28.48% / 12.17%

Table 2: Warm-up evaluation results. EM: exact match.

5 Experiments

We train for 3 epochs with per-device batch size 2 and gradient accumulation 16 on an RTX 4090 (24 GB). Evaluation metrics are CER (Track 1) and SER (Track 2) with sentence-level exact match for reference. We use seed 1337, learning rate 1×10^{-4} with 500 warmup steps, label smoothing 0.1, gradient checkpointing, TF32, and early stopping (patience 2). The HuggingFace Trainer default optimizer (AdamW) is used.

6 Results

On the warm-up set: Track 1 reaches 10.94% CER with 58.06% exact match; Track 2 reaches 28.48% SER with 12.17% exact match. These numbers are obtained with the shared pipelines and no external data beyond the provided corpora. We observe stable validation under balanced speaker splits and consistent normalization. Longer utterances show mildly higher error rates (notably in the 12.4–20 s bucket), and we observe small variations across DF/DM/ZF/ZM groups under the balanced-split protocol.

Final-test results. On the official final-test, our system achieves 18.78% CER for Track 1 (ranked 2/3 in our social group) and 33.38% SER for Track 2. For analysis, we also report a tone-removed Pinyin WER of 21.30% (ranked 2/2 among teams with available data). Figure 2 shows the official charts.

7 Reproducibility

Code and scripts are available at github.com/Kevindic0214/FSR-Challenge-

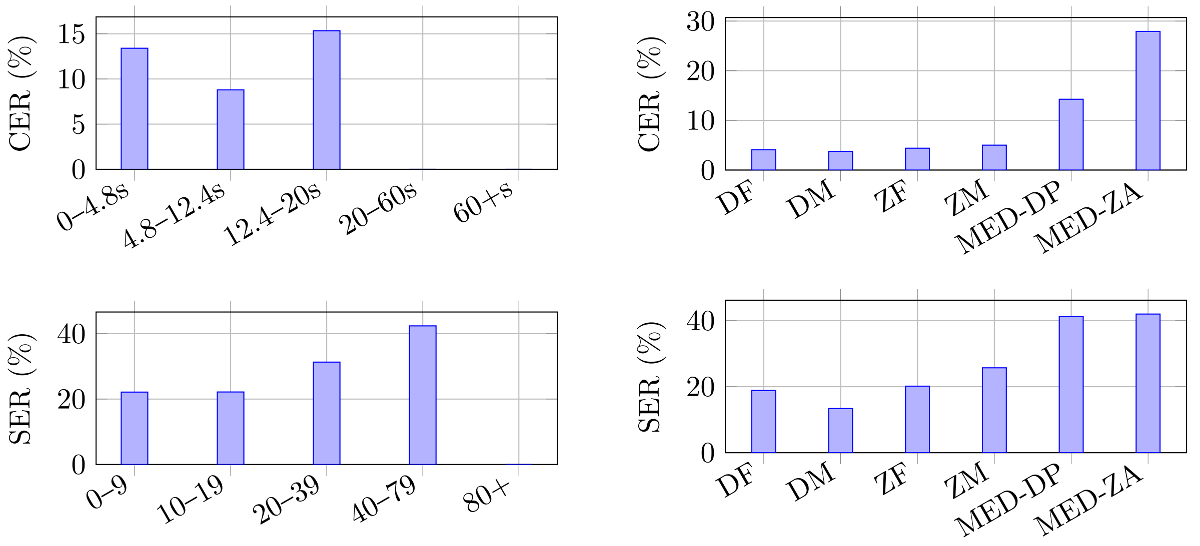


Figure 1: Warm-up analysis: (a) Track 1 CER by utterance duration; (b) Track 1 CER by group; (c) Track 2 SER by syllable length; (d) Track 2 SER by group.

Set	Track	Metric	Score (rank)
Final-test	1 (Hanzi)	CER	18.78% (2/3)
Final-test	2 (Pinyin)	SER	33.38% (2/2*)
Final-test	2 (Pinyin)	WER (tone-removed)	21.30% (2/2*)

Table 3: Official final-test results. *Among teams with available data in our social group.

2025.

We provide end-to-end scripts for data preparation, training, inference, and evaluation. Minimal examples and notes:

Notes. Track 1 defaults keep asterisks and punctuation unless `--strip_asterisk`/`--strip_punct` is specified. Track 2 drops starred syllables by default and supports optional split-tone merging with `--fix_split_tone`. All runs use seed 1337.

Track 1:

```
python prepare_hakka_track1.py --root
    HAT-Vol2 \
  --drop_mispronounce --relative_audio_path
python train_whisper_lora_track1.py \
  --train_jsonl
      HAT-Vol2/manifests_track1/train.jsonl \
  --dev_jsonl
      HAT-Vol2/manifests_track1/dev.jsonl
python infer_track1.py --eval_root
    FSR-2025-Hakka-evaluation \
  --outfile predictions_track1.csv --model
      openai/whisper-large-v2 \
  --lora_dir runs/track1/lora_v2_r16_e3
python eval_track1_cer.py --key_dir
    FSR-2025-Hakka-evaluation-key \
```

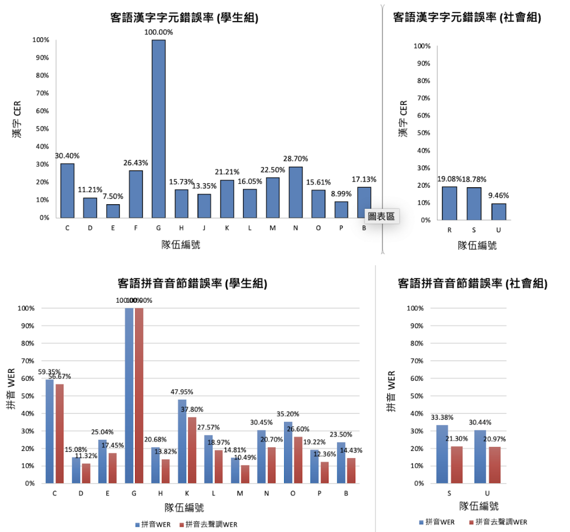


Figure 2: Challenge 2025 final-test result charts.

```
  --hyp predictions_track1.csv
```

Alternative (recommended for Track 2 with v3-turbo):

```
python train_whisper_lora_track2.py
    --base_model
    openai/whisper-large-v3-turbo --out_dir
    exp_track2_whisper_large_v3_turbo_lora
python infer_track2.py --eval_root
    FSR-2025-Hakka-evaluation \
  --outfile predictions_track2.csv --model
      openai/whisper-large-v3-turbo \
  --lora_dir
      exp_track2_whisper_large_v3_turbo_lora
```

Track 2:

```
python prepare_hakka_track2.py --data_root
    HAT-Vol2 \
```

```
  --out_dir HAT-Vol2/manifests_track2
      --exclude_mispronounced
python train_whisper_lora_track2.py
    --base_model
    openai/whisper-large-v3-turbo --out_dir
    exp_track2_whisper_large_v3_turbo_lora
python infer_track2.py --eval_root
    FSR-2025-Hakka-evaluation \
  --outfile predictions_track2.csv --model
      openai/whisper-large-v2 \
  --lora_dir exp_track2_whisper_large_lora
python eval_track2_ser.py --key_dir
    FSR-2025-Hakka-evaluation-key \
  --hyp predictions_track2.csv
```

8 Error Analysis

Common errors include character/phonetic substitutions and occasional short repeats; we monitor n-gram repetition to detect degeneration. Performance degrades mildly for longer utterances; bucketed analysis suggests length-aware decoding or better segmenting could help.

Examples. Sampled warm-up mismatches: *(003jh5p8hd.wav)* ref: 大家攏無仰子**啲**隨枷你人救出去; hyp: 大家攏無仰子**項**隨枷你研究出去.
(03qw9gfad7.wav) ref: 食**著**幾隻草蜢**乜好啊**; hyp: 食**到佢**隻草蜢**毋會好啊**.
(04qied7gz8.wav) ref: ...這兜**地動無幾著呢**莊頭...; hyp: ...這兜**地圖無幾臭呢啊**莊頭...

These illustrate homophone/near-neighbor substitutions and local phrase alterations; stronger language modeling or constrained decoding may mitigate such errors. For Pinyin (Track 2), common patterns include tone-digit confusions and occasional effects from star-syllable handling; our normalization reduces such artifacts.

9 Conclusion

We provide a concise, reproducible baseline for both tracks of FSR-2025 Hakka ASR using Whisper+LoRA. Future work includes dialect-aware adaptation, LM-rescoring for Hanzi, refined Pinyin normalization, and temperature/beam tuning.

Limitations

Our results are based on the provided HAT-Vol2 training data and the official warm-up set. We do not explore external language models or data augmentation; Pinyin normalization choices (e.g., starred syllables) can affect SER.

Acknowledgments

We thank the FSR-2025 organizers and dataset providers.

References

FSR2025. 2025. Formosa speech recognition challenge 2025: Hakka asr. Challenge. Warm-up evaluation set and official task description.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1, Dublin, Ireland. Association for Computational Linguistics.

HAT-Vol2. 2024. Hat-vol2: Taiwanese hakka speech corpus. Dataset. ~60 hours; Dapu and Zhao'an dialects; 16 kHz mono.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Optimizing Whisper Parameters and Training Data Processing for Formosa Speech Recognition Challenge 2025 - Hakka ASR II

李振皞 國立陽明交通大學資訊工程學系 howardlee.cs13@nycu.edu.tw
郭勝威 國立陽明交通大學資訊工程學系 guosw.cs13@nycu.edu.tw
鄭安喆 國立陽明交通大學資訊工程學系 andrewzheng.cs13@nycu.edu.tw
陳秉華 國立陽明交通大學資訊工程學系 binghua.cs13@nycu.edu.tw
劉逸安 國立陽明交通大學資訊工程學系 whitecat.cs13@nycu.edu.tw

Abstract

This paper presents the development and experimental process of our system for the Formosa Speech Recognition Challenge 2025 (Hakka ASR). The proposed system is built upon the OpenAI Whisper model. We achieved significant performance improvements for the Sixian dialect of Hakka through dataset preprocessing and model fine-tuning. In the warm-up evaluation, our system achieved a Character Error Rate (CER) of 10.51% on the character recognition track and a Syllable Error Rate (SER) of 14.72% on the pinyin recognition track. In the final evaluation, our system achieved a Character Error Rate (CER) of 11.21% on the character recognition track and a Syllable Error Rate (SER) of 15.08% on the pinyin recognition track.

Abstract

本文旨在記錄參與 2025 年福爾摩沙語音辨識挑戰賽（Hakka ASR）的實驗過程。我們所提出的系統以 OpenAI 的 Whisper 模型爲基礎，透過對官方提供的客語四縣腔資料集進行預處理與模型微調，顯著提升了辨識效能。在熱身賽的評估中，本系統於漢字辨識任務達到 10.51% 的字元錯誤率（CER），並在拼音辨識任務中達到 14.72% 的音節錯誤率（SER）；而在決賽評估中，本系統於漢字辨識任務達到 11.21% 的字元錯誤率（CER），並在拼音辨識任務中達到 15.08% 的音節錯誤率（SER）。

Keywords: Whisper, Hakka, Denoise

1 Introduction

1.1 深度學習與語音技術的發展

隨著深度學習 (deep learning) 的快速發展，語音相關技術在近幾年有許多突破，特別是在語音辨識 (Automatic Speech Recognition, ASR)，深度神經網路的應用大幅提升模型的準確性，現在的 end-to-end Transformer 模型逐漸取代傳統 ASR 系統。傳統模型通常拆成多個部分分開訓練在進行整合；相較之下 end-to-end 模型直接能從輸入的聲音訊號學習對應的文字輸出，簡化了訓練流程，降低了系統設計的複雜性，同時也增加了辨識準確度。

在許多 end-to-end 模型當中，近年來最具代表性的就是 OpenAI 所發表的 Whisper-ASR (1) 其系統從網路上蒐集眾多資料訓練而成。本篇研究選擇以 Whisper-ASR 作爲基礎模型，並在此基礎上進行針對客語語料的微調以及評估。同時，透過一系列實驗，如在資料加上 SpecAugment(2) 資料增強技術，以及嘗試應用 LoRA(3) 進行參數調整，以期待進一步提升模型的辨識準確度。

1.2 客語在台灣的發展

客語在台灣隨著漢語的普及，逐漸趨向弱勢，在語言傳承方面，現除了國民教育體系中的本土語言課程以外，透過逐漸發展的深度學習語音辨識來建立資料庫，甚至進一步結合相關技術提供學習資源，可做爲客語傳承的重要方法。

1.3 Formosa Speech Recognition Challenge 2025

此競賽聚焦於台灣客語語音辨識的相關研究，藉由提供客語語音轉拼音文字資料集，提供語音辨識模型開發的相關資源。比賽分爲兩部分，Track 1 以字元錯誤率 (CER) 來評估漢字輸出；Track 2 以音節錯誤率 (SER) 評估拼音輸出。透過這兩種評估指標，有效比較不同模型在漢字與拼音輸出上的辨識效果。

2 Methodology

2.1 Dataset

本研究所使用的資料主要來自兩部分。第一部分是競賽主辦方提供的官方訓練集，第二部分則是熱身賽的資料。我們將熱身賽資料全部納

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 471–475
November 20-22, 2025 ©2025 Association for Computational Linguistics

入最終訓練，以最大化資料利用率。詳細的資料統計如表 1 所示。總計約 72 小時的語料，涵蓋了多樣的說話者與內容，爲模型訓練提供了穩固的基礎。

Table 1: 訓練資料集統計 (Statistics of the Training Dataset)

Data Source	Duration (hrs)	Utterances
Train Set	62.02	27,349
Warm-up (Speech)	8.01	3,458
Warm-up (Media)	2.22	946
Total	**72.25**	**31,753**

2.2 Evaluation Metric

我們遵循比賽規定，採用字元錯誤率 (Character Error Rate) 與音節錯誤率 (Syllable Error Rate)，兩者公式如下：

$$ErrorRate = \frac{S+D+I}{N}$$

S 代表被替換的字元 (音節)，D 代表被刪除的字元 (音節)，I 代表被插入的字元 (音節)，而 N 代表總字元 (音節) 數。錯誤率越低，表示模型效能越佳。

2.3 Fine-tuning Whisper

儘管 Whisper 模型在多種語言上表現優異，但對於客語等特定低資源語言或領域，其詞彙覆蓋與聲學特徵適應性仍有提升空間。因此，我們採用微調（Fine-tuning）策略，在我們的客語資料集上進一步訓練預訓練好的 Whisper 模型。此舉能使模型學習客語獨特的發音、詞彙與語法結構，從而有效降低辨識錯誤率。

2.4 Data Augmentation

爲了解決資料量有限可能導致的過擬合問題，並提升模型的泛化能力，我們探索了多種資料增強（Data Augmentation）技術，模擬眞實世界的語音變化。

2.4.1 SpecAugment

SpecAugment (2) 是一種在時頻譜（Spectrogram）上進行遮蔽的有效增強技術。我們主要採用其中兩種策略：

- **Time Masking**: 在時間軸上隨機遮蔽一小段連續的訊框，模擬語音中短暫的停頓或遮蔽。
- **Frequency Masking**: 在頻率軸上隨機遮蔽一段連續的頻帶，增強模型對部分頻率資訊損失的魯棒性 (Robustness)。

2.4.2 Audio Concatenation

觀察到訓練集中單個語音檔案的平均長度較短，我們設計了語音拼接策略（Audio Concatenation），以期待讓模型更好地適應長語音輸入。具體作法爲：隨機選取數個（本實驗設爲 3 個）短音檔，將其拼接成一個新的、更長的音檔，其對應的標註也相應拼接。

2.4.3 Speed Perturbation

爲了模擬不同說話者的語速差異，我們對音檔進行速度微擾（Speed Perturbation）。透過改變音訊的採樣率來實現加速（如 1.1 倍）或減速，但不改變其音高。

2.4.4 Noise Injection

原始訓練集的錄音環境相對純淨。爲了提升模型在眞實噪音環境下的表現，我們在部分音檔中加入了高斯白噪音（Gaussian Noise），噪音的強度根據原始訊號的振幅進行設定。

2.5 LoRA

隨著模型規模的增大，完整的微調（Full fine-tuning）對計算資源的需求也急劇增加。爲此，我們嘗試了低秩適應（Low-Rank Adaptation, LoRA）(4) 技術。其核心思想是凍結預訓練模型原有的權重 W_0，並在模型特定層（如 Transformer 的 attention 層）旁注入一個可訓練的低秩矩陣 $\Delta W = AB$。原始的前向傳播 $h = W_0x$ 被修改爲：

$$h = W_0x + \Delta Wx = W_0x + ABx$$

其中，$W_0 \in R^{n\times m}$，$A \in R^{n\times r}$ 和 $B \in R^{r\times m}$ 是可訓練的低秩矩陣，且秩 $r \ll \min(n, m)$。如此，需要更新的參數數量從 $n \times m$ 大幅減少至 $(n + m) \times r$，從而顯著降低了訓練成本。

2.6 Denoising as Preprocessing

我們在決賽資料的語音部分含有噪音，而當我們用降噪工具進行處理後，發現即使原始的語音與降噪後的聽起來沒有太大差別，輸出也會不一樣。我們推論是因爲兩個音檔在轉成梅爾頻譜後會有明顯的差異。因此，我們決定將訓練集的資料也進行降噪以確保一致性。

3 Experiments

3.1 Experimental Setup

在本研究的所有實驗中，我們皆以 `whisper-medium` 作爲基礎模型（base model），並在 NVIDIA V100 GPU 上進行訓練。爲確保比較的公平性，所有模型的訓練週期（epoch）均設定爲 5 次，批次大小（batch size）爲 4，學習率（learning rate）則固定爲 5×10^{-5}。

我們使用的訓練資料集結合了主辦方提供的官方資料與熱身賽資料。爲了評估模型效能，我們將熱身賽資料集以 60% 與 40% 的比例進行切分，其中 40% 的部分作爲我們的測試集。所有實驗結果均在此測試集上進行評估，並以字元錯誤率 (Character Error Rate, CER) 與音節錯誤率（Syllable Error Rate, SER）作爲主要指標。

3.2 Effectiveness of Data Augmentation

爲了驗證不同資料增強技術對模型的影響，我們設計了一系列的對比實驗。

3.2.1 SpecAugment

首先，我們評估了 SpecAugment 的效果。實驗中，我們對每一筆輸入資料以 50% 的機率應用 SpecAugment，其中時間遮罩（Time Masking）的參數設爲 30，頻率遮罩（Frequency Masking）的參數設爲 15。如表 2及表 3 所示，僅加入 SpecAugment 就讓模型的 SER 從 9.51% 降至 9.04% 及讓 CER 從 3.58% 降至 3.45% ，顯示此技術能有效提升模型的泛化能力。

Table 2: SpecAugment 實驗結果比較 (拼音)

Configuration	SER (%)
Baseline (Original Data)	9.51
+ SpecAugment	**9.04**

Table 3: SpecAugment 實驗結果比較 (漢字)

Configuration	CER (%)
Baseline (Original Data)	3.58
+ SpecAugment	**3.45**

3.2.2 Other Augmentation Techniques

接下來，我們在 SpecAugment 的基礎上，進一步疊加其他三種增強方法：語音拼接（Audio Concatenation）、語速改變（Speed Perturbation）與噪音注入（Noise Injection）。

- 語音拼接: 隨機挑選 3 個音檔拼接成 5000 筆新資料。
- 語速改變: 將語速調整爲 1.1 倍。
- 噪音注入: 加入標準差爲 0.005 的高斯白噪音。

實驗結果如表 4 及表 5 所示。我們發現，語音拼接策略帶來了最佳效果，將 SER 與 CER 分別進一步降低至 8.96% 及 2.93%。然而，在拼音的部分，語速改變反而使模型表現略微下降，噪音注入的影響則相對中性。而在漢字部分，語速改變與噪音注入的表現皆些許上升。

Table 4: 多種資料增強技術實驗結果 (拼音)

Configuration	SER (%)
+ SpecAugment (Baseline)	9.04
+ SpecAugment + Audio Concatenation	**8.96**
+ SpecAugment + Speed Perturbation (1.1x)	9.61
+ SpecAugment + Noise Injection	9.06

Table 5: 多種資料增強技術實驗結果 (漢字)

Configuration	CER (%)
+ SpecAugment (Baseline)	3.45
+ SpecAugment + Audio Concatenation	**2.93**
+ SpecAugment + Speed Perturbation (1.1x)	3.18
+ SpecAugment + Noise Injection	3.12

3.3 Comparison of Base Models and LoRA

我們進一步比較了不同尺寸的 Whisper 模型，以及在大型模型上應用 LoRA 技術對效能的影響。本階段實驗包含以下設定：(i) 使用 `whisper-large-v2` 與 `whisper-large-v3-turbo` 進行微調；(ii) 在 `whisper-medium` 與 `whisper-large-v3` 上採用 LoRA 進行參數高效微調。爲加速實驗迭代，所有模型均僅訓練 3 個 epoch。各模型的 LoRA 參數（alpha, rank）如表 6 所示。

實驗結果顯示，`large-v2` 在客語資料集上的表現最佳（SER = **8.78%**），相較於 `large-v3-turbo`（SER = 9.95%）略有優勢。然而，僅透過 LoRA 微調的模型（`medium` 與 `large-v3`）效能仍顯著落後完整微調，顯示雖然 LoRA 能有效降低訓練參數量，但在本資料集上仍難以達到相同的準確度。

Table 6: 不同基礎模型與 LoRA 實驗結果 (拼音)

Model / Method	Alpha	Rank (r)	SER (%)
whisper-large-v3-turbo	–	–	9.95
whisper-large-v2	–	–	**8.78**
whisper-medium (LoRA)	128	256	22.24
whisper-large-v3 (LoRA)	128	256	21.80

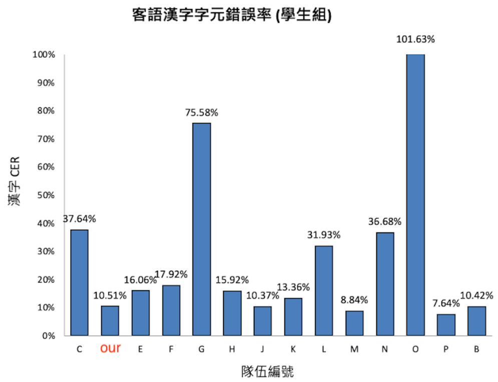
綜合以上所有實驗與時間成本的考量，我們最終決定採用以 `whisper-medium` 爲基礎，並僅使用 SpecAugment 作爲資料增強策略的模型，同時加入降噪處理以提升模型在含噪環境中的適應力。

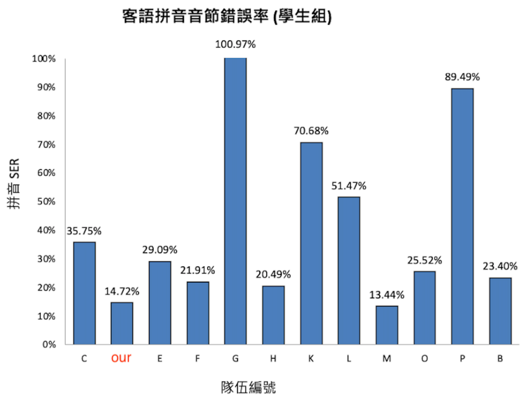
4 Conclusion

此篇研究以 Whisper-ASR 爲基礎，針對台灣客語的語音資料進行優化，透過資料前處理，SpecAugment 資料增強、語音拼接、語速調整、加入噪音以及 LoRA 技術等方法，嘗試提升模型的效能，由於目前我們所訓練的資料大多都是乾淨的聲音，所以這些資料增強無法明顯看出模型適應不同環境，但我們也自己設想，在什麼樣的環境下，哪種資料增強技術對最後的辨識結果會有較大的提升。熱身賽的成績如表 7所示。

Table 7: 熱身賽成績

Type	Character	Pinyin
CER(%)	10.51	X
SER(%)	X	14.72

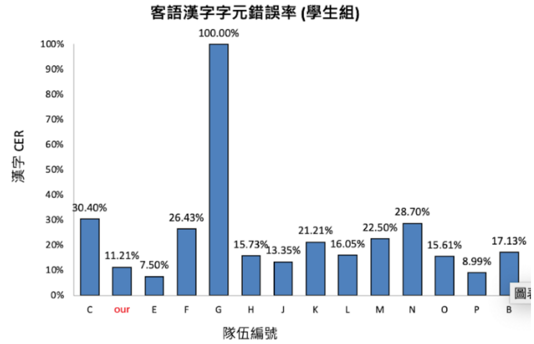


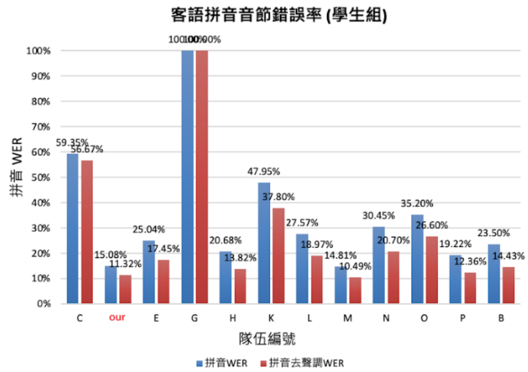


在決賽的測試資料中，我們發現語音檔案中含有大量雜音。爲了應對這個情況，我們採取的策略是：先將訓練資料進行降噪處理，再訓練一個新模型，接著也對主辦方提供的決賽資料進行同樣的降噪，最後再用訓練好的模型進行預測。決賽的成果如表 8 所示。

Table 8: 決賽成績

Type	Character	Pinyin
CER(%)	11.21	X
SER(%)	X	15.08
SER(去聲調)(%)	X	11.32





本研究初步展示了 Whisper 對於不同語言的可調整性，透過此開發以及實驗，提供了可行的方向，協助保存與推廣逐漸式微的台灣客語，未來可以朝向蒐集各種情境下的客語語料，來讓模型能夠適應各種不同的環境，以期在語言科技與文化傳承上做出更多貢獻。

Acknowledgment

感謝顏安孜教授在本研究過程中提供 GPU 資源，並感謝國家高速網路與計算中心提供環境讓我們使用，感謝 Formosa Speech Recognition Challenge 2025 主辦方提供訓練資料，讓本研究得以圓滿完成。

References

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint arXiv:2212.04356*, 2022.

[2] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A

Simple Data Augmentation Method for Automatic Speech Recognition," in *Proceedings of Interspeech 2019*, pp. 2613-2617, 2019. DOI: 10.21437/Interspeech.2019-2680.

[3] Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. *LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR.* arXiv preprint arXiv:2406.06619.

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. *LoRA: Low-Rank Adaptation of Large Language Models.* In ICLR.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is All you Need.* In NeurIPS.

The EZ-AI System for Formosa Speech Recognition Challenge 2025
針對 2025 福爾摩沙客語語音辨識競賽的 EZ-AI 辨識系統

Yu-Sheng Tsao[†], **Hung-Yang Sung**[‡], **An-Ci Peng**[‡], **Jhih-Rong Guo**[‡], **Tien-Hong Lo**[‡]
[†]EZ-AI, [‡]National Taiwan Normal University
sam@ez-ai.com.tw, redsheep@ntnu.edu.tw, anci.peng@gmail.com,
jhih.rong.guo@gmail.com, teinhonglo@ntnu.edu.tw

摘要

本研究參與 2025 客語語音辨識競賽的拼音軌及漢字軌，針對大埔腔與詔安腔兩個低資源腔調，設計並比較不同的語音辨識系統。我們的核心策略是透過跨語言遷移學習 (Transfer Learning)，有效利用相近語系的資源，並結合自監督學習 (Self-Supervised Learning, SSL) 以提升模型在拼音軌的辨識效能。在漢字軌方面，則使用 Whisper 模型並搭配 LoRA (Low-Rank Adaptation) 進行微調。爲了緩減語料不足的限制，我們採用兩種資料擴充方法：模擬對話式語音以處理多語者情境，以及利用文字轉語音 (Text-to-Speech, TTS) 生成額外的詔安腔語料。在熱身賽的結果顯示，遷移學習在拼音軌表現尤爲顯著，使系統於所有隊伍中取得平均字錯誤率 (Character Error Rate, CER) 19.57%，排名第三；在漢字軌中，Whisper 結合 LoRA 系統則達到平均 CER 6.84%，並獲得社會組第一名。本研究證明遷移學習與資料擴充能有效提升低資源語言的辨識表現，但在媒體語料的領域落差下仍存在挑戰，未來將探索語境學習 (In-Context Learning, ICL) 與熱詞建模 (Hotword Modeling) 以改善此問題。

Abstract

This study presents our system for Hakka Speech Recognition Challenge 2025. We designed and compared different systems for two low-resource dialects: Dapu and Zhaoan. On the Pinyin track, we gain boosts by leveraging cross-lingual transfer-learning from related languages and combining with self-supervised learning (SSL). For the Hanzi track, we employ pre-trained Whisper with Low-Rank Adaptation (LoRA) fine-tuning. To alleviate the low-resource issue, two data augmentation methods are experimented with: simulating conversational speech to handle multi-speaker scenarios, and generating additional corpus via text-to-speech (TTS). Results from the pilot test showed that transfer learning significantly improved performance in the Pinyin track, achieving an average character error rate (CER) of 19.57%, ranking third among all teams. While in the Hanzi track, the Whisper + LoRA system achieved an average CER of 6.84%, earning first place among all. This study demonstrates that transfer learning and data augmentation can effectively improve recognition performance for low-resource languages. However, the domain mismatch seen in the media test set remains a challenge. We plan to explore in-context learning (ICL) and hotword modeling in the future to better address this issue.

關鍵字：客家語、語音辨識、FSR-2025

Keywords: Hakka, Speech Recognition, FSR-2025

1 簡介

儘管語音辨識的研究於主流語種進展快速，欲解決的研究問題已從通用場景延伸至不同的小眾市場，但針對資源匱乏的弱勢語言，如何善加發揮仍然是一項具挑戰性的題目。

此次 2025 客語語音辨認競賽聚焦在兩個弱勢腔調：大埔腔、詔安腔。認知到這兩個腔調語料有限，欲達到較佳的辨識結果勢必會需要額外的語料參與，儘管腔調上有所不同，同屬客語語系的相似腔調語料在模式上還是會有所幫助 (Qian et al., 2024)。因此我們的做法分爲兩部分：最大化既有語料的表現，以及擴增目標語言的資料。

在低資源語言的研究領域中，使用預訓練模型是熱門且相對簡單的手段，如何能夠在有限資源中有效的學習也是此領域的一大重點 (Piñeiro-Martín et al., 2024)。方法上可以是：

- 尋找類似語系，經大規模語料訓練過的先進模型，對其解碼器進行遷移訓練

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 476–480
November 20-22, 2025

- 使用自監督模型如 wav2vec, WavLM 或 HuBERT 並訓練其進行下游的語音辨識任務 (Zhao and Zhang, 2022)
- 基於語音辨識基礎模型如 Whisper (Radford et al., 2023)，因其訓練所用的語料以及任務設定，使得 Whisper 能夠快速地適應不同的語料標記，並達到足夠強健的辨識結果這些方法讓稀少資源的語料也能善用既有的優異語音特徵，從有限的標記中達到較理想的辨識效果。

另外我們觀察到在測試語料中的媒體語料子集具有比較複雜的語音環境，如對話、噪音、遠場等性質，以及可能測試語料與訓練語料的領域差異導致表現不佳，我們參考過往研究與比賽經驗，合成相似性質的語料進行訓練，以改善辨識結果。

2 策略與方法

2.1 K2 與 SSL

K2 爲 Kaldi (Povey et al., 2011) 作者所建立團隊進行開發的語音處理框架，具辨識效果良好、易操作、運算快速、節省資源等優勢，並且在中文語系的常見語料都有預訓練模型可供快速實驗；另外在自監督模型上，也有對應的研究 (Yang et al., 2024) 能夠套用如 wav2vec、HuBERT 等模型，進行下游任務的訓練，故在漢字賽軌，我們會先已 HuBERT + RNN-T 的方式訓練，將資料集擴增的策略在此模型上做初步的嘗試。

在拼音賽軌，由於 Whisper 最相近於客語的語系爲中文，但解碼器在該語系已經被訓練至對漢字比較拿手，微調其輸出拼音，又或是重置解碼器都是相對次優的做法，所以拼音我們會使用 K2 zipformer 於 WenetSpeech 預訓練的模型，重置其解碼器使其輸出拼音。

2.2 Whisper

Whipser 爲 OpenAI 所發表的語音辨識基礎模型，使用常見的 Transformer 架構，訓練在 68 萬小時自網路蒐集、多數來自 Youtube 影片的多語言語料，訓練任務爲轉錄與轉譯（至英文）。由於語料的多樣性，Whisper 對於常見的環境變異都有良好的強健性，欲微調相似語系時也只需較少的語料就能有所改善，至今仍是熱門的語音辨識模型。

然而此模型若想要在中文上有可靠的辨識能力（準確度高於八成），至少得選用參數量 small 以上的模型，而訓練更大的模型卻伴隨著更長的訓練時間，不利於比賽的實驗迭代，故我們會先在 K2 探索適合的語料設定，再套用至強健性較佳的 Whisper，並使用 AdaLoRA (Zhang et al., 2023) 技術降低訓練的運算成本。

Corpus	Spks.	Sents.	hrs.
Train			
Dapu	64	12197	31.43
Zhaoan	59	15152	30.59
Eval (Pilot test)			
Studio - Dapu	10	1304	4.01
Studio - Zhaoan	11	2154	4.00
Media - Dapu	-	445	1.08
Media - Zhaoan	-	501	1.13
HakkaCouncil			
Reading - Sixian	208	-	396
Reading - Hailu	151	-	300
TTS - Zhaoan			
OOV	9	682	9.65
E-Learning	9	124588	136

Table 1: 比賽的訓練與測試語料，與 TTS 語料的統計資訊

Unique Words	Train	Eval (OOV)	
		Studio	Media
dapu	5771	2323 (230)	1552 (483)
zhaoan	4911	2221 (57)	1317 (275)

Table 2: 訓練與測試語料的詞目統計，括號中爲遺失字數量

2.3 資料擴增

在 K2 的初步實驗中，我們發現儘管對於錄音室測試語料的漢字辨識能力已能達到九成以上的正確率，在媒體測試語料上卻不到三成；同樣的情況也發生在 Whisper 的結果上，尤其是詔安腔的部分，與大埔腔的字錯誤率差了大約 2.5 倍。我們進一步分析發現，媒體語料相較於錄音室語料會有較多的遺失字 (Out-of-Vocabulary, OOV)，在詔安腔所以我們使用近期常見於稀少資源語料的擴增做法：透過 TTS 合成額外語料 (Chen et al., 2023)，來試圖提高在媒體語料上的表現。在這裡我們使用 FormoSpeech 團隊的 TTS 模型 yourtts-htia-240704[1] 進行語料的生成。

另外，由於媒體語料的組成大多爲對話式的語音，若只用朗讀型的語料訓練，模型在遇到語者的語音重疊或是被打斷時，辨識結果會產生明顯衰退，所以我們在 K2 的訓練額外合成了對話式的語料，以提高媒體語料的辨識率。

[1]https://huggingface.co/formospeech/yourtts-htia-240704

Exps. (CER%)	Studio			Media			Total Avg.
	Dapu	Zhaoan	Avg.	Dapu	Zhaoan	Avg.	
Train	6.78	6.46	6.62	73.98	80.01	77.00	41.81
+ft speed&reverb	6.15	5.64	5.90	68.48	77.77	73.13	39.51
Train&Conversation	6.12	5.25	5.69	63.71	74.92	69.32	37.50
Train&Conversation (TTS)	3.80	3.52	3.66	62.34	57.37	59.86	31.76

Table 3: 漢字軌於 k2 框架進行的實驗結果

Exps. (CER%)	Studio			Media			Total Avg.
	Dapu	Zhaoan	Avg.	Dapu	Zhaoan	Avg.	
FormoSpeech/hakka	9.47	29.95	19.71	14.56	41.19	27.88	23.79
+ft Train	1.11	2.22	1.67	8.71	21.13	14.92	8.29(6.84)
+ft Train & OOV (TTS)	1.13	3.26	2.20	7.85	21.11	14.48	8.34

Table 4: 漢字軌於 Whisper 的實驗結果，總平均欄位的括號爲官方所回報之結果

3 實驗設定

3.1 資料集

除了決賽的結果會加入熱身賽的測試語料進行訓練外，其他實驗的訓練語料皆不包含測試語料。這些語料的統計資料如表 1。

TTS 語料進一步分成兩種，我們先使用客語能力認證的文字語料進行一般性用詞的語音生成，但發現只用這個領域的文字語料並不足以改善媒體測試語料的辨識率，我們便仔細檢視媒體語料的標記，並與客家詔安腔字典進行比較，如表 2，鎖定分詞結果不在訓練語料的句子進行合成。雖然就統計上來看大埔腔在 OOV 的字詞比例較多，但由於詔安腔的表現較差，故我們優先以詔安腔進行語料的合成。

3.2 硬體與參數

在本次的大部分實驗中，我們採用兩款不同型號的 GPU 進行運算，分別爲 NVIDIA GeForce RTX 3090 與 4090。不論是在 k2 框架下或是使用 Whisper，所需時間均大約爲 18 至 24 小時。在 k2SSL 的實驗上我們參考原始論文的訓練參數，訓練最多 200 週期 (epoch)，再挑選收斂的區間進行權重平均後，使用貪婪搜尋法 (Greedy Search) 進行解碼。

對於 Whisper 的類型挑選，我們站在巨人的肩膀上，使用 FormoSpeech 團隊所公開的 whisper-large-v3-taiwanese-hakka[2] 模型作爲基底，此模型使用台灣最常見的六種客語腔調進行微調訓練，直接對這次比賽的測試語料辨識就已經具有不錯的表現，我們即固定使用這個模型做爲基準，進一步使用這次比賽語料進行微調，另外也參考了先前比賽的報告，加入資料擾動如：速度、音高變動與空氣吸收 (AirAbsorption) 以適應測試語料媒體子集的聲學環境，最多訓練 10 個週期。推論則挑選驗證集損失最低的單一檢查點，大部分收斂落在 3~5 週期左右。

4 實驗結果

4.1 漢字軌

4.1.1 K2SSL 實驗

初步實驗我們採用 K2SSL 研究中的 zipformer-based HuBERT 模型 (由 HuBERT-base-ls960 衍生) 作爲編碼器訓練 RNN-T 系統進行辨識，如表 3，僅使用比賽訓練資料的話，雖然能在錄音室語料上達到接近九成五的辨識率，但在會議語料上卻僅有兩成左右，即使進一步增加資料的擾動，改善的程度也有限。

觀察媒體語料的組成後，我們將訓練語料加上擾動，產生模擬對話情境的語料再次訓練。模擬對話情境相較僅使用朗讀風格的訓練語料有更爲明顯的改善，但在媒體語料的部分，詔安腔的表現則明顯弱於大埔腔。故我們蒐集詔安腔能力測驗的文字語料，使用 TTS 產生詔安腔的合成語料後，再次模擬對話情境進行訓練，在詔安腔媒體語料降低了 17% 的字錯誤率，並也一併改善了錄音室語料的辨識率。

然而媒體語料的整體辨識率仍不及五成，我們推測由於 HuBERT-base 因僅訓練在 Librispeech 的朗讀語料，仍不具有足夠的強健性處理複雜的聲學情境，因此接下來我們會使用 Whisper 進行。

4.1.2 Whisper

實驗結果如表4，我們將 FormoSpeech 團隊所微調的模型作爲基準值，使用其直接對測試

[2] https://huggingface.co/formospeech/whisper-large-v3-taiwanese-hakka

Exps. (CER%)	Studio			Media			Total Avg.
	Dapu	Zhaoan	Avg.	Dapu	Zhaoan	Avg.	
Zipformer-HuBERT	4.55	19.15	11.85	35.50	59.34	47.42	29.64
Wenet-Zipformer +ft '23, '25 train	5.77	10.78	8.69	20.20	40.29	31.37	17.17
Wenet-Zipformer +ft 客委會->'23, '25 train	5.46	10.36	8.31	21.07	38.75	30.90	16.76
Whisper +ft Train	8.32	17.24	12.78	26.56	31.52	29.04	20.91(19.60)

Table 5: 拼音軌的實驗結果，總平均欄 (Total Avg.) 的括號內數字爲官方所回報之結果

語料進行辨識，在大埔腔的錄音室與媒體語料均有接近九成的辨識率，得益於 Whisper 對複雜聲學環境的強健性，詔安腔則相對較爲弱勢，所以在 Whisper 的實驗上我們仍然是聚焦在改善詔安腔的辨識結果。

使用比賽的訓練語料進行微調後，在錄音室語料上就有大幅度的改善，兩個腔調的平均字錯誤率從 19.71% 下降至 1.67%，推測是基底模型在訓練時詔安腔語料不足的關係；媒體語料也從平均 27.88% 下降至 14.92%，儘管如此，媒體語料的詔安腔錯誤率仍居高不下，即使我們進一步針對分詞後的 OOV 去產生合成語料，也僅僅是讓大埔腔的辨識結果稍微改善，詔安腔的改善仍然有限。

對此，針對媒體語料進一步分析錯誤結果，應是媒體語料含有比例不少的專有名詞，導致即使模型已經在不同聲學環境、額外的 OOV 合成語料上訓練了，面對專有名詞依然是無法妥善的辨識。

4.2 拼音軌

我們使用在漢字軌上較爲有效的策略訓練拼音軌的模型：在 k2 框架上採用自監督模型或是預訓練模型，並適時增加語料，考慮到儘管腔調不同，拼音書寫均爲一致。在 Whisper 則是直接使用比賽語料進行訓練。

由於 Whisper 解碼器的設計，將其重置再訓練將會喪失訓練過大量語料的優勢，故我們沿用原本的設定，微調中文語言讓他能夠輸出拼音。而 k2 模型因沒有這類限制，所以我們能夠直接訓練其解碼器輸出拼音。

實驗結果如表5，在錄音室語料上，兩種 k2 模型的拼音辨識結果都比 Whisper 更加準確；在媒體語料方面，即使 Whisper 因預訓練語料，比起 Zipformer-HuBERT 表現更穩定，但其優勢並不如漢字軌一般明顯，一旦換上 WenetSpeech 預訓練過的 Zipformer (下稱 Wenet-Zipformer)，只需針對拼音解碼的模型在整體的辨識效果上仍比較理想。如果我們兩階段的先將 Wenet-Zipformer 用客委會[3] 的資料微調，再微調至 2023 & 2025 年的比賽資料，能進一步改善模型的辨識結果，在熱身賽的測試資料上達到平均字錯誤率 16.76%。

4.3 熱身賽結果

因爲熱身賽的時程關係，繳交的時候我們在漢字與拼音軌均使用 Whisper 的結果進行投稿，漢字軌錯誤率 6.84% 取得了社會組及所有隊伍的第一名，而拼音軌則取得了錯誤率 19.57%，位居所有隊伍的第三名。

4.4 決賽結果

考慮到決賽的語音可能也會與媒體測試語料相似，我們將 75% 的媒體語料加入漢字軌 Whisper 的訓練，訓練過程的評估指標則使用錄音語料與剩下的媒體語料計算，加入部分媒體語料後的測試集可以觀察到明顯的改善，若使用決賽語料去評估加入媒體語料前後的辨識結果之字元差異，也能得到 10% 左右的差異結果，故我們使用這顆 75% 媒體語料的模型進行漢字軌辨識的結果提交，得到 CER 9.46% 的成績，位居所有隊伍的第三名。拼音軌我們使用 Wenet-Zipformer 進行提交，儘管拼音軌應能直接的辨識出不同腔調的拼音序列，但訓練時並無納入媒體測試語料以及客委會媒體語料，或許導致模型在複雜聲學環境仍不夠強健，最終拿到了拼音 WER 30.44% 的成績，位於所有隊伍的第八名。

5 結論與展望

此次比賽我們參考過去的實驗結果與經驗，透過分析標記並增加合成語料，試圖改善詔安腔媒體語料存在過多遺失字與專有名詞，導致漢字軌辨識率居高不下的情況，不過因爲增加的絕大多數都屬於領域外資料，改善有限。未來我們會試著探討語境學習 (Incontext Learning) 或是熱詞等方式進行擴增或調校，改善領域外資料的辨識效果。

[3]https://www.aclclp.org.tw/doc/hat_brief_c.pdf

References

Po-Kai Chen, Bing-Jhih Huang, Chi-Tao Chen, Hsin-Min Wang, and Jia-Ching Wang. 2023. Enhancing automatic speech recognition performance through multi-speaker text-to-speech. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 371–376, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docio-Fernandez, María del Carmen López-Pérez, and Georg Rehm. 2024. Weighted cross-entropy for low-resource languages in multilingual speech recognition. In *Interspeech 2024*, page 1235—1239. ISCA.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, et al. 2011. The kaldi speech recognition toolkit.

Mengjie Qian, Siyuan Tang, Rao Ma, Kate M. Knill, and Mark J.F. Gales. 2024. Learn and Don't Forget: Adding a New Language to ASR Foundation Models. In *Interspeech 2024*, pages 2544–2548.

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.

Yifan Yang, Jianheng Zhuo, Zengrui Jin, Ziyang Ma, Xiaoyu Yang, Zengwei Yao, Liyong Guo, Wei Kang, Fangjun Kuang, Long Lin, et al. 2024. k2ssl: A faster and better framework for self-supervised speech representation learning. *arXiv preprint arXiv:2411.17100*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

Jing Zhao and Wei-Qiang Zhang. 2022. Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1227–1241.

多模組錯誤檢測與修正的客語語音辨識系統
A Multi-Module Error Detection and Correction System for Hakka ASR

Min-Chun Hu	Yu-Lin Xiao	Wen-Hsiang Lu
National Cheng Kung University Department of Computer Science and Engineering	National Cheng Kung University Department of Computer Science and Engineering	National Cheng Kung University Department of Computer Science and Engineering
p76131563@gs.ncku.edu.tw	p76121518@gs.ncku.edu.tw	whlu@mail.ncku.edu.tw

摘要

本研究提出一個針對客語（以大埔／詔安腔為主）的自動語音辨識（ASR）後矯正系統，旨在解決低資源語言辨識錯誤率偏高的問題。客語因受限於語料規模、異體字與腔調差異，在既有的通用 ASR 模型上表現往往不佳。為此，我們首先以 Whisper Large v3 Turbo 為基底辨識模型，使用約 60 小時的大埔與詔安語料進行微調，以提升對特定腔調的適應性。在獲取 ASR N-best 候選句後，系統進一步透過多模組錯誤偵測矯正流程進行修正，包含四個主要步驟： (1) 潛在錯誤偵測，用於鎖定候選間錯誤的候選詞彙；(2) 音素混淆集偵測（Phoneme Confusion Set）: 依據音素相近關係提供可能替代詞；(3) 辭典（Lexicon）修正: 確保詞彙存在於語言使用的實際範疇中，(4) 搭配詞關聯度偵測: 利用收集之語料所建立的搭配詞關聯度來偵測錯誤詞彙。本研究所提出的矯正機制能有效補足 ASR 在低資源語言中的不足，實驗顯示經過多階段錯誤偵測矯正後，最終 CER 減少至 15.49%，減少 2.14 % ，證明該方法能有效提升語音辨識的準確率。

關鍵字：語音辨識、客語、錯誤矯正、混淆集、搭配詞

。

Abstract

This study proposes a post-correction system for Automatic Speech Recognition (ASR) targeting Hakka (with a focus on the Dapu and Zhao’an dialects), aiming to address the high error rate commonly observed in low-resource languages. Due to limitations in corpus size, the existence of variant characters, and dialectal differences, Hakka often performs poorly on general-purpose ASR models. To improve recognition, we first fine-tuned Whisper Large v3 Turbo with approximately 60 hours of Dapu and Zhao’an speech data, enhancing the model’s adaptability to these specific dialects. After generating the ASR N-best candidates, the system performs a multi-module error detection and correction process consisting of four main steps: (1) potential error detection to identify suspicious words among candidates; (2) phoneme confusion set detection, which provides alternative words based on phonetic similarity; (3) lexicon-based correction to ensure that words belong to valid linguistic usage; and (4) collocation-based detection, which leverages word association scores derived from collected corpora to identify contextually inconsistent words. The proposed correction mechanism effectively compensates for the limitations of ASR in low-resource languages. Experimental results show that, after multi-stage error detection and correction, the final Character Error Rate (CER) was reduced to 15.49%, achieving a 2.14% absolute reduction, thereby demonstrating that the method can effectively enhance ASR accuracy.

Keywords: speech recognition, Hakka, error correction, confusion set, collocation

1 緒論

客語屬於低資源語言，現有公開語料相對稀缺，且漢字用法中存在異體字、詞彙多形

以及方言差異等挑戰。在自動語音辨識（ASR）的輸出結果中，常見的錯誤類型包括：（i）同音近形所造成的用字錯誤、（ii）華客語彼此搶詞引發的混淆，以及（iii）語意搭配不精確等問題。這些錯誤不僅影響辨識結果的可讀性，也限制了 ASR 系統在真實場域中的應用效益。因此，本文旨在 ASR 輸出後進行精細化的矯正，以提升整體客語辨識的準確度與實用性。

本研究所提出的 ASR 系統基於 Whisper Large v3 Turbo 的基底辨識模型，並以約 60 小時的大埔與詔安腔語料進行 fine-tuning 訓練。在獲得 N-best 候選句後，系統會依序經過多階段的錯誤偵測矯正模組：首先透過跨候選句的詞彙差異統計來標記潛在詞彙錯誤位置；其次利用混淆集提供音素近似的替換建議；再透過搭配詞分數檢查語境合理性；最後以辭典檢驗輸出，避免輸出未收錄或極罕見的詞彙。此流程兼顧可擴充性與可解釋性，適合於商業應用實務中逐步迭代改進。

然而，本研究同時也面臨低資源語言的典型挑戰：一方面，深度學習模型如 Whisper 在大規模語料上能展現優異的表現，但在客語這類語料有限的情境下，辨識效能往往因缺乏詞彙完整覆蓋性而受到限制；另一方面，若僅依靠人工擴充語料，則需付出龐大的人力與時間成本，而異體字及方言差異更使資料標註的一致性難以維持。因此，如何在「模型效能」強化與「語料資源不足」改善之間取得平衡，並透過多模組的後處理錯誤偵測矯正機制來彌補 ASR 的不足，即是本研究的核心挑戰與主要貢獻。

2 相關研究

2.1 語音辨識模型（Speech Recognition Model）

傳統的音素式（phoneme-based）ASR 模型（Daniel et al., 2011）在錯誤矯正上具有一項天然優勢：它們會產生音素層級的輸出，可作為辨識後矯正（post-recognition correction）的額外參照資訊。相較之下，非音素式（non-phoneme-based）End-to-End model（如 Baevski et al., 2020；Radford et al., 2023）多半是直接從語音映射到文字，缺少顯式的音素資訊。像 Kaldi 與傳統混合式 ASR 這類音素式系統，會保留音素序列與對齊資訊，因而能支援基於音素的對齊與矯正策略。然而，儘管音素式方法在錯誤矯正上具可用訊息，其整體辨識正確率通常仍不及最新的 End-to-End model。這類系統高度依賴發音詞典、音素對齊與人工詞彙資源，在跨語言或跨領域時彈性受限。在近年的 ASR model 中，Whisper（Radford et al., 2023）以其在多語言、多領域上的穩健性與廣泛實務採用特別受到關注；其在大規模有標註的平行語料上訓練，使其具備良好的泛化能力，在未特別微調的情況下亦能有不錯表現。不過，在特定領域或是特定語言(例如:客語的特殊腔調等等)，Whisper 的表現仍可能不理想。主要原因在於預訓練語料中領域相關資料稀缺，導致辨識錯誤偏多。因此，實務上常需要額外的調適或錯誤矯正機制，才能在此類高風險情境中達到可靠表現。

2.2 非音素式 ASR 的錯誤矯正（Non-phoneme-based ASR Error Correction）

在 ASR 後矯正領域，非音素式的方法亦有顯著進展。許多研究（Ma et al., 2023）不再僅依賴單一最佳假設（1-best），而是利用 N-best 假設作為輸入，以提供較豐富的候選與語境訊息，提升矯正的準確率。儘管如此，面向特定領域的 ASR 矯正仍具挑戰，尤其在專業術語與語言變體的處理上。部分研究（López-Cózar & Callejas, 2008）嘗試將語意、句法、詞彙與語境納入模型，使矯正更貼近自然語言使用情境，進而得到更精準且自然的修正。

2.3 音素式 ASR 的錯誤矯正（Phoneme-based ASR Error Correction）

近年亦有越來越多工作將音素層級訊息用於改善 ASR 後矯正。不同於傳統必須倚賴辭典與對齊的音素式 ASR，一些方法改為在辨識之後再運用音素序列來偵測與修補可能錯誤。例如，Serai et al.（2019）提出以聲學模型的後驗機率來進行抽樣，取代固定的混淆矩陣，藉此較真實地模擬 ASR 錯誤，並使錯誤行為更貼近現代 ASR 的實際狀況。另有方法（Wang et al., 2022）將音素（phonetic）與語意（semantic）資訊結合 N-best 假設共同

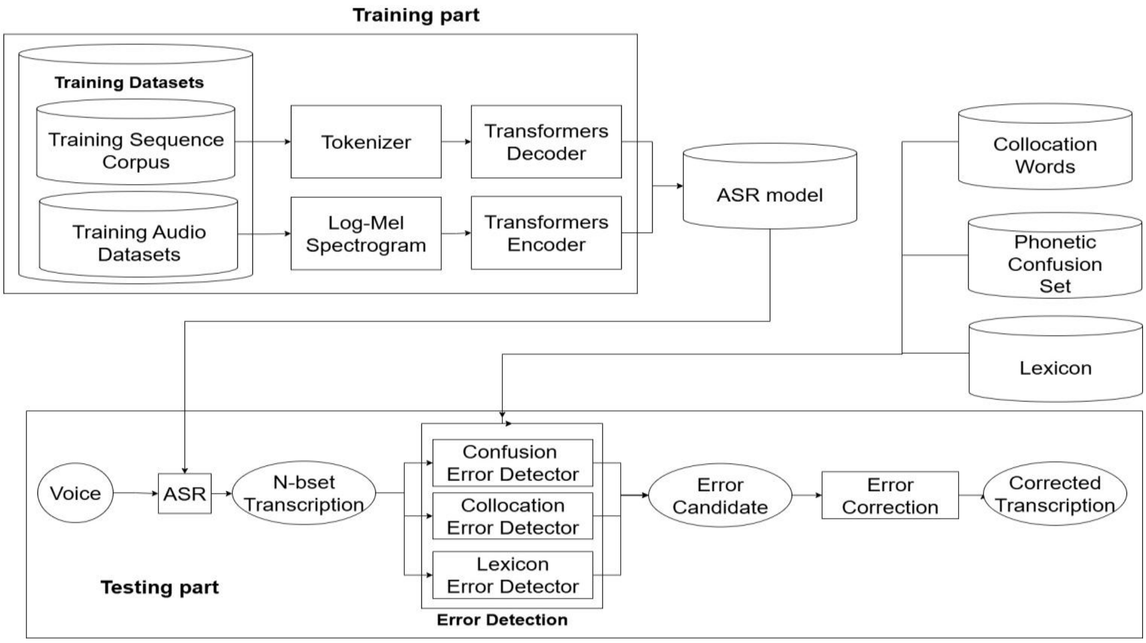


圖 1:系統架構

使用，用以偵測「聽起來合理但語意不通」的錯誤，進一步提升矯正準確率。

3 研究方法

3.1 系統架構

- Training part

在訓練階段，本研究以客語（大埔／詔安）語料為基礎，建構資源較稀少的語音辨識模型。首先語音資料被轉換為 Log-Mel Spectrogram，文字資料則經過 Tokenizer 轉為符號序列，兩者分別進入 Transformer Encoder 與 Transformer Decoder 進行訓練(參考圖 1)。透過編碼器與解碼器的互動，完成語音與文字的對應學習，最終得到基於 Whisper Large v3-turbo 微調之 ASR model。此模型能夠捕捉客語聲學與語言特徵，為後續測試與錯誤檢測模組奠定基礎。

- Testing part

本研究設計之客語（大埔／詔安）語音辨識矯正系統包含三大特徵(參考圖 1)，分別為自動語音辨識（ASR）、潛在錯誤偵測、混淆集合（Confusion Set）、語意搭配檢測（Collocation）、辭典比對（Lexicon）以及最終輸出。整體流程如圖 1 所示。

在 ASR 階段，系統以 Whisper large v3-turbo 為核心，並利用 60 小時的大埔與詔安語料進行微調。於解碼過程中，模型會產生多個候選句輸出（N-best），提供後續模組進一步比對與修正。

接下來的錯誤偵測模組透過候選之間的對齊與統計分析，辨識在相同位置上出現多種不同字詞的情況，並將其標記為潛在錯誤點。其內部包含三個子模組：(1) 音素混淆偵測模組(2) 辭典比對模組(3) 語意搭配檢測模組。

最後，錯誤候選會進入 Error Correction 模組。該模組依據錯誤檢測所提供的候選字詞，結合辭典與語境約束，挑選出最合適的修正結果，最終生成 Corrected Transcription，提供較為穩健且符合語意的客語逐字稿輸出。

3.2 語音前後處理

在進行語音辨識前，系統會先進行前處理。首先，利用 Silero-VAD 偵測並擷取有效的語音區段，去除靜音與雜訊部分，確保輸入訊號的品質。接著，所有音訊皆統一重取樣為 16 kHz 並轉換為單聲道，以符合模型的輸入需求。此外，輸出候選詞之文字亦需正規化處理，包含將簡體字轉換為繁體字，以及針對異體字進行客語漢字替換，以避免因文字形式不一致而導致後續比對失敗。

3.3 潛在錯誤區域偵測

潛在錯誤偵測模組以漢字為最小對齊單位，針對多個候選輸出進行逐字對齊，並統計每個位置的輸出詞彙多樣性。若在相同位置上出現兩種以上的不同結果，即判定為潛在錯誤區域。此設計利用了 N-best 解碼輸出在錯誤區域往往呈現高分歧度的特性，藉由辨識候選間的差異，能夠有效捕捉可能的錯誤字詞，並作為後續修正模組的輸入依據。接著系統將透過三個錯誤偵測模組來進行處理: (1)音素混淆偵測模組:根據音近或形近關係，建立潛在替代詞的候選集，(2) 辭典比對模組:整合客語辭典作為修正依據(3) 語意搭配檢測:透過大規模文本計算詞與詞之間的相關性，評估候選詞或片語與句內其他詞語之間的搭配合理性。

3.4 錯誤檢測

錯誤檢測模組的核心在於辨識 ASR 輸出中潛在的異常字詞，作為後續矯正的基礎。系統透過多候選序列的比對與對齊，觀察在同一位置出現高分歧度的情況，並將其標記為可疑區段。此方法能有效捕捉音素混淆、語境不符或辭典外詞彙等錯誤來源，縮小矯正範圍並提升整體修正的精準性。

3.4.1 音素混淆偵測模組

本模組中，系統同時考慮字級與短語級的混淆情境，以建立潛在的替代候選。於字級層面，系統根據客語的聲母、韻母、鼻化、送氣與腔口等音素差異，為每一潛在錯誤字生成音近的候選字詞。例如，聲母之間的變異如 s/ts/tsh、p/ph/f 以及 n/l 等，皆是常見的聲母混淆來源。於短語層面，則考慮整體語境下的音近短語，當某些詞組在語音或語意上高度相似時，即納入候選。例如，「貧」與「鼻」在拼音上為 pin113 與 pi53，在實際辨識過程中容易因發音相近而產生混淆。此時，音素混淆集合（Phoneme Confusion Set）(參考表 1)便能提供可能的替代字詞，協助系統在後續的偵測與修正階段進行判斷。

漢字及音素	說明
貧↔鼻 pin113↔pi53	拼音相近，常在口語中混淆
呼↔福 fu113↔fug21	拼音相近，常在口語中混淆
過↔擱 go53↔gog21	拼音相近，常在口語中混淆

表 1 : Phoneme confusion error 範例

3.4.2 錯誤詞彙偵測模組

在錯誤詞彙偵測階段，系統藉由客語辭典進一步過濾與修正候選詞。若原始辨識候選詞不在辭典中，但其混淆候選詞存在於辭典內，則系統傾向將輸出修正為候選詞。舉例來說(參考表 2)，當系統將「頒獎」誤辨識為「班獎」時，由於「班獎」並非辭典中的詞彙，而「頒獎」則是存在於辭典中，系統便會修正為「頒獎」，當作矯正候選。此外，為確保修正結果的合理性，系統遵循「從左至右、最長片語優先」的原則，並避免修正片段之間的互相重疊。此外，系統也能新增詞彙，例如，當輸入為「新聞高」時，辭典中並無此詞，但其音近的「新聞稿」是合理詞彙。此時將「新聞稿」新增至辭典中，並將辨識結果修正為「新聞稿」，以便後續辨識與修正能更準確地處理此類詞彙。

原詞 -> 修正	理由
新聞高 → 新聞稿	辭典無”新聞高”，但近似音”新聞稿”存在辭典中
讀立 → 獨立	辭典無”讀立”，但近似音”獨立”存在辭典中
班獎 → 頒獎	辭典無”班獎”，但近似音”頒獎”存在辭典中

表 2 : 辭典修正

3.4.3 語意搭配詞檢測

在大埔與詔安腔的口語逐字稿中，部分辨識錯誤並非單純由於語音近似或句法不合所導致，而是源自於詞語搭配上的語境不合理。例如，若辨識結果為「蠶窟的事實」，則顯得語意不通，因為「蠶窟」與「事實」並非合理組合；相對地，「殘酷」與「事實」則經常並置使用，語境上更為自然。同樣地，若辨識結果為「學曉天光日當晝會辦一場考

試」也會顯得語意不通順，因為「學曉」與「考試」的無搭配詞關聯度，但「學校」往往與「考試」共同出現，形成高頻搭配，進而將「學曉」替換成「學校」。為了系統化偵測此類語境層面的不一致，本研究在偵測流程中引入搭配詞錯誤檢測模組。

此模組採用多個通用主題語料庫，包括教育、醫療、公共／新聞以及日常生活等領域，並透過目前蒐集到的客語文本（如新聞稿、廣播逐字稿、教材與口語語料）計算雙詞之間的搭配詞關聯度。其核心概念是比較兩個詞在語料中同時出現的頻率，與它們各自單獨出現的頻率進行對比，用以衡量兩詞的相關性。分數越高，代表兩個詞在語境中越常被搭配使用，也越符合語境搭配。

例如，「殘酷」與「事實」通常具有較高的搭配詞關聯度，因此為自然的詞語組合；相反地，「蠶窟」與「事實」的關聯度明顯偏低，語意上難以成立，系統會傾向將「蠶窟」修正為「殘酷」。同樣地，「學校」與「考試」的搭配頻繁度很高，因此在候選包含「學曉／學校」的情況下，系統會選擇語境上更合理的「學校」。

3.5 錯誤矯正

完成錯誤偵測後，系統進入錯誤矯正階段。此階段的重點是整合各模組所提供的候選詞，依據語音特徵、詞彙合法性與語境搭配等面向進行比對與篩選，輸出最符合語言使用情境的結果。矯正並非單一規則，而是多種訊息的綜合判斷。例如：

- 音素混淆集合：若將「貧民窟」誤辨為「鼻民窟」，可依音近關係修正，恢復合理詞彙。
- 辭典比對：當輸出為「該係吾夢相」時，透過辭典判定後修正為「該係吾夢想」，以回復正確語義。
- 在搭配詞的情境下，若辨識結果出現「學曉考試」，因「學曉」與「考試」缺乏語境關聯，而「學校」與「考試」則是高頻搭配，系統便能將「學曉」修正為「學校」。

透過這些不同來源的訊息整合，錯誤矯正模組能顯著提升逐字稿的自然度與準確性，為最終輸出提供更穩健的保證。

4 實驗

本研究的 ASR 模型基於 OpenAI Whisper large-v3-turbo，並在客語大埔與詔安語料上進行微調，所得到的模型命名為 Hakka_dapu_za 在微調過程中，輸入採用 16 kHz 的 mel-spectrogram，解碼則使用 top-k 與 top-p 採樣策略以生成 N-best 候選序列，同時透過 Silero VAD 去除靜音與雜訊，確保輸入訊號的品質。為了驗證本研究方法的有效性，我們同時設置了 baseline 進行比較：baseline 模型為 Whisper large-v3-turbo，未經針對客語語料進行額外調整。藉由與 baseline 的對照，我們可以清楚評估微調與錯誤矯正模組對於辨識準確率的實際貢獻。

4.1 資料集

在語種與腔調方面，本研究聚焦於客語大埔腔與詔安腔。訓練語料主要採用客語競賽所提供的約 60 小時大埔與詔安語音資料（涵蓋日常對話、新聞播讀、教學講述等），並以此對 OpenAI Whisper large-v3-turbo 進行微調。語料同時涵蓋兩種腔調，並盡量在性別、年齡及錄音條件（錄音室／半自然環境）上保持平衡。所有音檔在前處理階段均被轉換為單聲道 16 kHz，以符合模型輸入需求，此外，本研究在錯誤偵測與修正模組中，分別建立三種輔助資源，拼音字典、辭典比對、以及搭配詞關聯度：

(1)拼音字典:根據語料所附的拼音與漢字，建立客語漢字對羅馬拼音的映射字典。此資源用於將客語漢字逐字轉換為拼音，並支援大埔與詔安常見用字。若遇到未收錄的字則標記為 NULL。

(2) 客語辭典:整合教育部客語辭典、大埔腔辭典與詔安腔辭典，建立詞條對應拼音的資料庫，避免輸出極少見的詞彙。

(3) 搭配詞關聯度計算:利用蒐集之文本計算詞與詞之間的關聯度，衡量語境中的搭配合理性。資料來源包含教育、醫療、公共／新聞及日常生活等領域的文本（涵蓋客語辭典、哈客平台文章、客語朗讀材料等，共計 16,324 篇以客語漢字撰寫的文章）。矯正時不限於相鄰詞，系統會在全句上下文中進行雙向比對（A→B 與 B→A）。

4.2 初始模型比較與選擇

在本研究的實驗設計中，第一步需要確定最適合作為後續研究基礎的初始模型。我們針對此部分進行了兩種設定的比較。在實際測試時，我們將初始的 60 小時語料進行分割，其中 50 小時作為訓練集，10 小時作為測試集。第一種設定是以 Whisper Large v3 Turbo 模型為基底，並先在約 800 小時的「海陸四縣」語料上進行微調。這樣的模型理論上在跨腔調任務上可能具備一定的泛化能力。然而實驗結果顯示，在此設定下模型的 CER（字錯誤率）為 16.91%，表現並不理想。第二種設定則是直接使用 Whisper Large v3 Turbo 的預訓練版本，在相同測試條件下其 CER 為 15.46%，較優於前者。

表格中 M1 為 Whisper Large v3 Turbo 先以海陸／四縣語料做基底，再以本研究 50 小時 大埔＋詔安語料微調。M2 為 Whisper Large v3 Turbo 預訓練版直接以本研究 50 小時 大埔＋詔安語料微調（不混入其他腔調）。

	M1	M2
CER	16.91%	15.46%

表 3: 模型準確率比較

根據表 3 可見，當在初始模型中混入不同腔調的語料（如海陸四縣）時，會造成負遷移效應，進而影響辨識精度，反而對目標腔調（大埔與詔安）的辨識任務產生負面影響。因此，本研究最終選擇不納入其他腔調的語料，而是直接以 Whisper Large v3 Turbo 的預訓練版本作為後續微調的基底模型，並專注於大埔與詔安的語料，期望藉由更聚焦的語言特性來獲得更高的辨識準確率。
在確定使用 M2 模型當作後續語音辨識模型。在此設定下，模型於客語競賽提供的熱身賽語料上表現的 CER 為 17.63% 。與未經微調的模型相比已提升 2.3%，但辨識結果中仍然存在部分錯誤，特別是來自音素相近的混淆、語境搭配不當以及異體字詞的使用等。這些問題若僅依賴單純的 ASR 模型仍不易解決，因此我們將本研究提出的後校正流程套用於模型的輸出，以進一步提升準確度與可讀性。本小節的實驗結果說明了初始模型選擇的重要性。避免跨腔調語料的干擾是提升精度的關鍵，而專注於大埔與詔安語料的設定，則為後續矯正模組的發揮提供了最佳的基礎。

4.3 矯正效果與分析

在確定基底模型後，本研究將完整的三階段矯正流程應用於 ASR 輸出，並選用 2025 年客語語音競賽所釋出的熱身賽語料進行測試，該資料集共計約 10 小時，涵蓋大埔與詔安腔的多樣語音內容，包括日常對話、新聞播報等等，能有效模擬真實使用情境。經過矯正處理後，系統的最終字 CER 從 17.63%降低至 15.49%，整體提升幅度為 2.14%。這樣的結果清楚顯示，本研究設計的多模組矯正流程，能有效修正音素混淆、語境搭配不當，以及辭典外詞彙的錯誤，並在低資源語言環境中展現顯著的實用價值。

表格中的 M2 代表未經矯正模組處理後之模型，M3 代表經矯正模組處理之模型。

	M2	M3
CER	17.63%	15.49%

表 4: 矯正模組前後之模型準確率比較

從具體例子(如表 5)來看，矯正模組能成功修正如「學曉 → 學校」這類音近字詞混淆，並利用考試相關語境進行合理替換；又如「讀立 → 獨立」，雖然音同，但「讀立」並不存在於辭典中，因此系統最終將其修正為合法的「獨立」，提升了輸出結果的可讀性與正確性。

原始辨識結果(M2)	修正後結果(M3)	原因
學曉天光日當晝會辦一場考試	學校天光日當晝會辦一場考試	“學曉”和”學校”音近可用搭配詞根據考試進行矯正
厥等在世界个盡頭過讀立个生活	厥等在世界个盡頭過獨立个生活	獨立 音同讀立，但”讀立”不在辭典內，將它修正成”獨立”

表 5: 矯正範例說明

4.4 討論

總體而言，實驗結果呈現兩個重點：第一，在初始模型的選擇上，避免將其他腔調語料納入訓練是必要的，因為這樣能降低跨腔調干擾帶來的負遷移；第二，在模型的基礎上，再結合我們提出的矯正流程，可以顯著改善模型的準確率，使其在客語（大埔、詔安）的辨識任務中展現更穩健的性能。

以下針對三類錯誤分析說明:

(1) 字詞過短或缺乏上下文的詞彙：(參考表 6)如「教師」「教育」等單詞或短語，測試音檔過短等的問題。這類詞通常需要更長的句子上下文（如搭配「課綱／考試／授課／學校」等）才有足夠訊號做決策。

(2)字詞消失:(參考表 6)模型直接遺漏了某些應有的詞彙，使輸出序列缺少關鍵的語義成分。與一般的替換錯誤或音近詞混淆不同，字詞消失並非來自候選詞之間的錯配，而是序列生成本身的缺陷。這種情況在基於 Transformer 的架構中特別常見，可能導致部分詞彙在輸出時被忽略。由於後端的矯正模組主要依賴「候選對照」與「上下文推斷」，一旦關鍵詞未被輸出，就無法建立映射關係，自然也無法補回遺漏詞語。因此，字詞消失問題通常需要透過前端 ASR 模型的改進來解決，例如增強聲學建模、調整解碼策略或引入更強的語言模型，而非單純依靠後端矯正流程。

(3) 專有名詞辨識錯誤：(參考表 6)屬於難以矯正的情境。人名，地名，組織名等等屬於專有詞彙，通常不在語言模型或辭典的高頻詞範疇內，加上聲韻結構多樣（如壽氏麗），容易被誤辨為發音近似或隨機組合的詞（如受勢力）。由於候選詞缺乏正確對應，加上人名在語境中往往缺少語意輔助，因此後端矯正難以將錯誤修復為正確的人名。

辨識結果	標準答案	原因
教師	教授	語境過短單詞修正較困難
厥等夢想使得食晝	行到半爛燦厥等夢想使得食晝	字詞消失無法補回遺漏詞
十四歲个沙百裡	十四歲个沙伯利	特殊人名錯誤缺乏候選詞對應

表 6 : 無法矯正範例

5 結論

總結來說，本研究針對客語（大埔與詔安腔為主）語音辨識的準確率提升，提出了一套結合 N-best 潛在錯誤偵測、Confusion Set、Collocation 與 Lexicon 的多階段錯誤矯正方法，實驗顯示經過多階段錯誤矯正後最終 CER 減少至 15.49%，減少 2.14 % ，證明該方法能顯著提升語音辨識的準確率與可用性，並為低資源語言的 ASR 矯正研究提供了一個具體且可行的解決方案。

未來研究方向主要著重於語言資源的擴充與優化。目前系統所依賴的辭典與搭配詞庫雖已涵蓋一般日常語境，但在專業領域（如醫療、教育、公共服務等）仍存在不足，導致在處理專業詞彙或專門語境時，系統的穩定性與準確率可能受到限制。若能進一步蒐集並整合專業語料，持續擴充辭典與搭配詞庫，將能有效提升矯正模組對於專業詞彙的辨識與修正能力，進而提升模型可靠性與泛化能力。

References

Povey, D., Ghoshal, A., Boulianne, G., et al. (2011). The Kaldi speech recognition toolkit. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011).

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (NeurIPS 2020).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023).

Serai, P., Wang, P., & Fosler-Lussier, E. (2019). Improving speech recognition error prediction for modern and off-the-shelf speech recognizers. In Proceedings of IEEE ICASSP 2019.

Guo, J., Wang, M., Qiao, X., Wei, D., Shang, H., Li, Z., Yu, Z., Li, Y., Su, C., Zhang, M., Tao, S., & Yang, H. (2023). UCorrect: An unsupervised framework for automatic speech recognition error correction. In Proceedings of IEEE ICASSP 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS 2017).

Yeh, C.-F., & Lee, L.-S. (2015). An improved framework for recognizing highly imbalanced bilingual code-switched lectures with cross-language acoustic modeling and frame-level language identification. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Wei, V. J., Wang, W., Jiang, D., Song, Y., & Wang, L. (2024). ASR-EC benchmark: Evaluating large language models on Chinese ASR error correction. arXiv preprint arXiv:2412.03075.

A Whisper-Based System with Multi-Faceted Data Augmentation for Low-Resource Language

Pin-Cheng Chen*[1] Yu-Chi Chen*[1] Chia-Chun Liang*[1]
Cheng-Yu Lin*[1] Ping-Juei Tsai*[1] Wei-Yun Ma†[1]

[1]**Institute of Information Science, Academia Sinica, Taipei, Taiwan**

b10102102@ntu.edu.tw, chi.mg10@nycu.edu.tw, d14948004@ntu.edu.tw
gary2004930518@gmail.com, allcare.c@nycu.edu.tw, ma@iis.sinica.edu.tw

Abstract

This paper presents a comprehensive approach for the Formosa Speech Recognition Challenge 2025 (FSR-2025), targeting automatic speech recognition (ASR) for the under-resourced Dapu and Zhao'an dialects of Taiwanese Hakka. Our method integrates data augmentation and robustness techniques, including SpecAugment, dialect-aware special tokens, text-to-speech (TTS) augmentation, noise/reverberation mixing, and speed perturbation, to mitigate data scarcity and domain mismatch. Experiments on the official FSR-2025 datasets show consistent improvements in both character error rate (CER) and word error rate (WER). Extensive ablation studies further confirm that each component contributes positively. These results offer a practical path toward robust ASR for under-resourced Hakka dialects and suggest broader applicability to other low-resource languages.

Keywords: Automatic Speech Recognition, Data Augmentation, Low-resource, Taiwanese Hakka

1 Introduction

Automatic Speech Recognition (ASR) has made remarkable progress in recent years, driven by large-scale speech corpora and powerful deep learning models. End-to-end pipelines have become standard, where Connectionist Temporal Classification (CTC)-based models provide efficient monotonic alignment (Graves et al., 2006), and attention/Transformer frameworks enhance long-range modeling (Chan et al., 2016; Dong et al., 2018; Barrault et al., 2023). For scenarios with limited labeled data, self-supervised pretraining (e.g., wav2vec, HuBERT) yields substantial gains by learning robust acoustic representations from unlabeled audio (Schneider et al., 2019; Baevski et al., 2020; Hsu et al., 2021). Building on this foundation, large models like Whisper leverage multilingual, multitask training to generalize across diverse languages and domains (Radford et al., 2022). More recently, Multimodal Large Language Models (MLLMs) have extended this paradigm by processing speech and text within a unified framework, facilitating cross-modal reasoning and transfer (Rubenstein et al., 2023; Zhang et al., 2023).

However, these advances are unevenly distributed: most data and models target high-resource languages (e.g., English, Mandarin), whereas minority and dialectal languages remain underserved. Taiwanese Hakka is a low-resource Sinitic language with multiple dialects; among them, Dapu and Zhao'an are particularly under-resourced. These challenges make it difficult for standard ASR models to achieve satisfactory performance, creating a significant technological gap for their speakers. The Formosa Speech Recognition Challenge 2025 (FSR-2025) directly addresses this issue by providing a benchmark dataset to foster research in this area.

To address these challenges, our methodology centers on a multifaceted data augmentation strategy, combining SpecAugment, text-to-speech (TTS) synthesis, noise and reverberation mixing, and speed perturbation. We also introduce dialect-aware special tokens to guide the model in distinguishing between the Dapu and Zhao'an dialects. The effectiveness of this approach and the contribution of each component are systematically evaluated through a se-

*Equal contribution. Authors are listed in alphabetical order by last name, except the advisor.
†Advisor and corresponding author.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 489–498
November 20-22, 2025 ©2025 Association for Computational Linguistics

ries of experiments and ablation studies on the official dataset, as detailed in the subsequent sections.

2 Dataset and Task Definition

Our study is based on the HAT-Vol-2 corpus, provided by the organizers. The corpus contains roughly **70** hours of audio from approximately **100** speakers across Taiwan and is divided into **three** official splits: *train*, *evaluation*, and *final-release* (test set).

The FSR-2025 challenge structure the task into two parallel tracks, each corresponding to a different orthography. This dual-track system defines the output targets for our models and the metrics for evaluation.

2.1 Orthography and Evaluation Tracks

- **Track 1: Recommended Hakka Characters.** This track uses a set of Han characters promoted by the Taiwanese Ministry of Education for writing Hakka. While leveraging semantic context familiar to readers of Sinitic languages, these characters often lack a one-to-one phonetic correspondence. For instance, the character '行' can have multiple pronunciations depending on the context. Performance on this track is measured by **Character Error Rate (CER)**.

- **Track 2: Hakka Pinyin System.** This track employs a phonemic transcription system that precisely represents initials, finals, and tones. It clearly distinguishes dialectal variations (e.g., the word "person" (人) is transcribed as `ngin113` in Dapu vs. `ngin53` in Zhao'an). However, this system is unfamiliar to most native speakers. Performance on this track is measured by **Word Error Rate (WER)**, where each Pinyin syllable is treated as a word.

In addition to the official data, we employ VoxHakka, a multi-accent, multi-speaker text-to-speech (TTS) system (Chen et al., 2024)[1], to synthesize additional Hakka speech. This mitigates data scarcity and expands both lexical and speaker coverage; our generation policy and settings are detailed in Section 4.3.

[1] https://voxhakka.github.io/

3 Model

3.1 Whisper

The Whisper model, developed by OpenAI (Radford et al., 2022), is an end-to-end ASR system with strong multilingual performance. Our study builds upon the **`whisper-large-v3-taiwanese-hakka`** checkpoint (hakka-whisper) (FormoSpeech, 2025), already fine-tuned on six Hakka dialects, and we further fine-tuned it on Dapu and Zhao'an data for FSR-2025. It's quite notable that the further evaluation revealed a divergence between character-based (Track 1) and pinyin-based (Track 2) transcription: the adapted checkpoint improved character recognition, but the original Whisper model performed better on pinyin, likely due to its broader phonetic generalization. This also highlights a trade-off between dialect adaptation and phonetic robustness.

3.2 MLLM-based model

In addition, we evaluate LLM-based approaches for speech transcription. Specifically, we use Kimi-Audio, which is based on the Qwen architecture, as the backbone of the language model (KimiTeam et al., 2025). Kimi-Audio employs a 12.5 Hz audio tokenizer and has been trained on large-scale Chinese text and audio data; it shows strong performance on Mandarin ASR benchmarks—indicating robust capability for Sinitic phonetic and orthographic patterns. This setup allows us to probe how well a large Chinese-trained LLM can transfer its knowledge to low-resource dialects such as Dapu and Zhao'an Hakka, and whether the model can effectively leverage its linguistic knowledge to compensate for the scarcity of labeled speech data.

4 Metholodgy

Due to the different effects on each track, we applied different methods to each of them. The utilized results are summarized in Table 1.

Table 1: Methods for Track 1 & 2

Track	SpecAugment	Special token	TTS
Track 1	✓	✓	✓
Track 2		✓	✓

4.1 SpecAugment

SpecAugment is a simple yet effective method that operates directly on the log-mel spectrogram (Park et al., 2019). Instead of relying on additional data, it improves model robustness toward noise by applying several types of transformations: time warping, frequency masking, and time masking. Time warping distorts the spectrogram along the temporal axis, while frequency masking and time masking randomly remove consecutive frequency channels or time steps, respectively. In our setting , we adopted frequency masking and time masking with a progressive enhancement strategy during training (Li et al., 2022; Lu and Li, 2024), which is also applied in images with good performance (Jarca et al., 2024).

4.2 Special token

In Whisper, special tokens can be utilized to control specific attributes of the speech recognition process such as task type, target language, and timestamping behavior (Radford et al., 2022). In practice, the token serves as a high-level cue for the model, guiding the model's acoustic and lexical predictions. Recent studies demonstrate that leveraging special tokens, which is often termed *prompt-based control*, can significantly improve Whisper's performance, particularly in low-resource or unseen language scenarios (Peng and Yan, 2023; Yang et al., 2024; Huang et al., 2025). For instance, studies have shown that introducing explicit prompts, such as language-family tags or even learnable soft prompts, helps guide the model toward more accurate transcriptions for underrepresented languages (Yang et al., 2025). Inspired by these findings, our work investigates a similar approach by introducing dialect-aware special tokens. We hypothesize that providing an explicit cue to distinguish between the closely related Dapu and Zhao'an dialects will enable the model to better activate dialect-specific acoustic and linguistic knowledge, thereby improving recognition accuracy for both.

4.3 Text-to-Speech

We synthesize additional Hakka speech with VoxHakka, a YourTTS-based, multi-speaker, multi-dialect TTS system for Taiwanese Hakka (Chen et al., 2024). We adopt a twofold generation policy with external and internal sources. On top of that, each transcript is generated in **three** voices, sampled randomly from VoxHakka's multi-speaker bank.

External sources *External* denotes text not included in the official data transcripts. We collect sentences from the Ministry of Education Hakka Dictionary[2] and the online teaching materials released by the Hakka Affairs Council (HAC)[3]. Given Han-character inputs, VoxHakka synthesizes the corresponding waveforms and generates pinyin labels, which are not provided by these sources. The synthesized utterances enrich the training set with terms and sentences that are rarely observed in spontaneous speech.

Internal sources *Internal* denotes text derived from official data transcriptions. We employ two strategies:

1. **Tokenized rare-term augmentation** We observed that official evaluation set often contains proper nouns and other low-frequency words that are scarce in the training set, making them a common source of recognition errors. To mitigate this, we first identify these rare lexical items from the training transcripts using a GPT-4o model (Hurst et al., 2024) guided by a carefully designed few-shot prompt. After de-duplication, each unique term is synthesized into an audio clip using the VoxHakka TTS system. This provides the ASR model with explicit acoustic examples of rare and potentially out-of-vocabulary (OOV) terms.

2. **Voice conversion** The released training set contains many repeated prompts recorded by multiple speakers–some sentences are read by up to 14 speakers–while

[2] https://hakkadict.moe.edu.tw
[3] https://elearning.hakka.gov.tw

other sentences occur only once or twice. Motivated by prior findings that Voice conversion (VC) based speaker augmentation improves ASR in low-resource settings (Baas and Kamper, 2021), we apply VC to under-covered sentences to increase speaker diversity: each such sentence is uttered by at least three distinct speakers.

5 Experiments

For evaluation, we adopted the official scoring mechanism provided by the competition[4]. Specifically:

- **Track 1:** Character Error Rate (CER) was used as the primary metric.
- **Track 2:** Word Error Rate (WER) was used as the primary metric.

5.1 Models

Our initial experiments focused on the Kimi-Audio model. It was fine-tuned on the *FSR-2025-train* set and evaluated on the *FSR-2025-evaluation* set. The system obtained a CER of 51.87% on Track 1 and a WER of 89.49% on Track 2. Notably, the outputs contained several abnormal generation artifacts[5]; manually correcting for these reduced the CER to 33.47%.

We then evaluated Kimi-Audio-Instruct, a variant trained predominantly on Mandarin data (KimiTeam et al., 2025), under the same configuration. This model yielded a CER of 45.70% on Track 1, which improved to 28.27% after correcting for the same abnormal outputs. For comparison, a hakka-whisper baseline trained with an identical setup achieved a markedly lower CER of 7.64%. Thus, while Kimi-Audio-Instruct outperformed the original Kimi-Audio, both models remained significantly behind the specialized hakka-whisper system.

Furthermore, our error analysis of both Kimi-Audio models revealed a particular weakness in processing longer utterances and tokens rare in the training data (e.g., proper nouns and transliterated names). These conditions not only yielded substantially higher error rates but also occasionally triggered the generative artifacts noted above, severely degrading overall performance.

To investigate the effect of data distribution, we conducted a controlled experiment. We created a new data split by merging the *FSR-2025-train* and *FSR-2025-evaluation* sets. From this combined pool, we held out 20% as a new development set and randomly sampled 5,000 utterances for a test set. Under this controlled setting, Kimi-audio achieved a greatly improved performance of **6.13% CER** (Track 1) and **7.56% WER** (Track 2). This result, summarized in Table 2, indicates that the model's performance improves dramatically when the evaluation data distribution is well-represented in its training data—especially concerning rare words and proper nouns.

Despite this promising result, we prioritized the Whisper-based system for the final challenge submission. This decision was based on the observed instability (i.e., the generation of abnormal outputs) and the higher computational cost of the Kimi-based models, which posed practical risks when facing an unknown final test set. Nevertheless, our findings suggest that MLLM-based approaches like Kimi-Audio hold considerable promise for future work, provided sufficient data coverage and improved model stability.

Table 2: Performance of Kimi-audio under a controlled split (train+evaluation merged; 20% held out; 5,000-item test sample).

Model	CER (Track 1)	WER (Track 2)
Kimi-Audio	6.13%	7.56%

5.2 Methods

5.2.1 Evaluation Setup

After confirming the model, to evaluate the usability of the proposed methods for ASR tasks, we adopted the following data split for testing on a Whisper-like model:

- **Training data:** 90% of (*FSR-2025-train* + 80% of *FSR-2025-evaluation*).
- **Validation data:** 10% of (*FSR-2025-train* + 80% of *FSR-2025-evaluation*).

[4] https://github.com/yfliao/FSR-2023-Hakka-ASR-Scoring

[5] For examples, see https://github.com/MoonshotAI/Kimi-Audio/issues/101

- **Testing data:** 20% of *FSR-2025-evaluation.*

We trained a *hakka-whisper* model on the training data, which serves as the **baseline** for comparison against the results of the different methods.

5.2.2 SpecAugment

In our setting, we set the following corrected progressive strategy: in the early stage, SpecAugment was applied with a probability of 30%, which means each batch would have a 30% chance of being perturbed using a time mask of 40 frames and a frequency mask of 14 bins. In the middle stage, the probability was increased to 50% with masking parameters set to 60 frames for time masking and 20 bins for frequency masking. Finally, in the late stage, the probability was set to 70% , with stronger augmentation using 80 frames for time masking and 27 bins for frequency masking.

The 30% initial probability (rather than 10%) provides sufficient augmentation from the start to prevent early overfitting, while the 70% final probability (rather than 80%) avoids over-augmentation that could harm model convergence. This balanced progression aligns with curriculum learning principles where moderate difficulty increases lead to better generalization (Jarca et al., 2024).

The progressive augmentation probability at step t is defined as:

$$p(t) = \begin{cases} 0.3 & \text{if } t/t_{max} < 0.3 \\ 0.5 & \text{if } 0.3 \leq t/t_{max} < 0.7 \\ 0.7 & \text{if } t/t_{max} \geq 0.7 \end{cases}$$

where t_{max} represents the total training steps.

At first, we compared the baseline to the one with noise mask. We conducted this test on Track 1. Surprisingly, the baseline achieved a CER of 4.47%, while applying the proposed method reduced the CER to 3.77% at 5000-step training. This corresponds to a relative reduction of 15.66% in CER, indicating a substantial improvement.

Secondly, we evaluated the effect of the progressive and the stationary enhancement. This time we conduct the experiment on Track 2. As shown in Fig. 1, we observed that under the stationary setup, the error rate plateaued

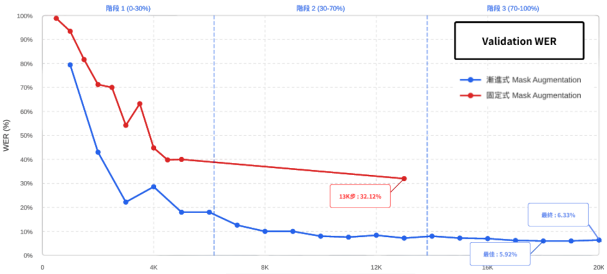


Figure 1: Comparison of validation WER between stationary (red-line) and progressive (blue-line) enhancement setups on Track 2. We utilized *Mask* Augmentation. Under the stationary setup, the error rate plateaued after around 4,000 steps and remained at about 32% by 13,000 steps. In contrast, the progressive setup continued to improve throughout training, reaching a validation WER of 5.92% at 20,000 steps (final value: 6.35%).

after approximately 4,000 steps. By 13,000 steps, the validation WER was still around 32%. In contrast, the progressive setup continued to improve throughout training. After 20,000 steps, the validation WER reached 5.92% (final value: 6.35%).

On the test set, the stationary setup yielded a WER of 46.53%, while the progressive setup achieved a significantly lower WER of 21.59%. This result validates our 30%-50%-70% progressive schedule, where the moderate initial augmentation (30% vs. 10%) allows faster convergence while the conservative final stage (70% vs. 80%) maintains stability.

5.2.3 Source-Aware Speed Perturbation

Unlike traditional uniform speed augmentation, we implement a source-aware speed distribution strategy that adapts to the inherent characteristics of different audio sources in the FSR-2025 dataset—which we identified based on the accompanying metadata—as shown in Table 3. Our analysis reveals significant heterogeneity: media sources (broadcasts, podcasts, etc.) exhibit fast speaking rates (1.1-1.3x relative to conversational speech) with high audio quality, while recorded conversational data show slower rates (0.8-1.0x) but suffers from reverberation and environmental noise.

Design rationale for asymmetric speed distributions:

- **Media sources** are heavily biased to-

Table 3: Source-aware speed factor distribution across different types of sources. The table summarizes the relative proportions of playback speed factors observed for *Media Source*, *Recorded Source*, and *General Source*. Overall, Media Source tends to concentrate in the slower range (0.70–1.00x) with a peak at 0.75x, while Recorded Source shifts toward faster speeds (0.90–1.15x) peaking at 1.00x. General Source covers a wider range (0.85–1.20x) and peaks at 1.05x.

Speed Factor	**Media Source**	**Recorded Source**	**General Source**
0.70x	15%	-	-
0.75x	25%	-	-
0.80x	20%	-	-
0.85x	15%	-	5%
0.90x	10%	10%	10%
0.95x	10%	15%	15%
1.00x	5%	20%	20%
1.05x	-	25%	20%
1.10x	-	20%	15%
1.15x	-	10%	10%
1.20x	-	-	5%
Range	0.70-1.00	0.90-1.15	0.85-1.20
Peak	0.75x (25%)	1.00x (20%)	1.05x (20%)

wards slowdown factors (75% probability in 0.70-0.85x range) to compensate for their inherently fast speaking rate. This prevents the model from overfitting to rapid speech patterns that are rare in target applications.

- **Recorded sources** receive balanced bidirectional augmentation with a peak at 1.0x (20%) and symmetric distribution (1.00-1.15x speedup, 0.90-0.95x slowdown). This addresses the slower baseline rate while maintaining natural variation.

- **General sources** adopt the widest range (0.85-1.20x) with a slight speedup bias (peak at 1.05x, 20%), maximizing robustness to diverse speaking rates in unknown data.

The progressive speed augmentation schedule follows three distinct phases (see Table 4), synchronized with SpecAugment and noise augmentation to achieve curriculum learning effects.

Rationale for progressive probability schedule: The middle stage employs the highest augmentation probability (0.6) as the model has developed sufficient acoustic modeling capacity to benefit from aggressive data perturbation while avoiding early-stage confusion. The late stage deliberately reduces augmentation intensity (0.4) to prevent over-regularization that could harm fine-grained learning of tonal patterns, critical for Hakka's complex tone system. Noise augmentation follows a particularly conservative schedule ($0.2 \rightarrow 0.5 \rightarrow 0.3$) because excessive noise can disrupt the *fundamental frequency* (F0) contour information essential for tone discrimination in Hakka.

Table 4: Progressive learning schedule across different training phases. The early stage (0–30% epochs) adopts conservative settings for foundation learning, the middle stage (30–70%) uses maximum augmentation for robustness building, and the late stage (70–100%) reduces augmentation to stabilize convergence.

Parameter	**Early Stage**	**Middle Stage**	**Late Stage**
Objective	Warm-up	Intensive	Stabilization
Epoch Range	0–30%	30–70%	70–100%
SpecAug Prob	0.3	0.6	0.4
Speed Prob	0.3	0.6	0.4
Noise Prob	0.2	0.5	0.3
Mask Intensity	0.7x	1.0x	1.2x

This coordinated multi-type augmentation strategy, validated through 30% relative CER reduction compared to uniform augmentation (from 4.47% to 3.13%), demonstrates the effectiveness of curriculum-based training for low-resource ASR.

Next, as shown in Table 5 with Track 1, we want to know the effects from the three masks: spectrum deformation (Spec), noise addition (Noise), and speed variation (Speed).

Overall, adding Speed augmentation leads

Table 5: CER results for different augmentation settings across training steps.

Setting	Stage	Step	CER (%)
Baseline	-	20k	4.33
Spec+Noise	Early (30%)	3k	4.02
Spec+Noise	Early (30%)	4k	3.91
Spec+Noise	Mid (50%)	5k	3.47
Speed+Spec+Noise	Early (30%)	3k	3.71
Speed+Spec+Noise	Mid (50%)	6k	3.40
Speed+Spec+Noise	Late (70%)	12k	**3.13**

to a consistent decrease in CER as the training steps increase. The best result, achieved at 12000 steps with Speed+Spec+Noise, shows an improvement of approximately 27.7% over the baseline (from 4.33% down to 3.13%).

5.3 Special tokens

We design an enhanced dialect conditioning mechanism by injecting dialect-specific tokens into the decoding process:

Dialect Token Insertion & Detection. We define a set of dialect tokens: <| 大埔腔 |>, <| 詔安腔 |>, and <| 未知腔 |>. Since the training datasets are labeled with the corresponding dialect, our system detects the dialect of each audio file based on rule-based metadata during preprocessing. And the system would prepends the appropriate token to the transcription text.

Balanced Sampling. To prevent majority dialects from dominating training, we employ a balanced sampling strategy. Depending on the configuration, batches can be drawn either equally from each dialect (equal strategy), or weighted to favor minority dialects.

We integrated both dialect tokens and balanced training and the results are presented in Table 6 and Table 7.

Table 6: CER results on Track 1 for with and without the Special token setting.

Setting	Step	CER (%)
Baseline	20000	4.33
Special token	30000	**3.48**

5.4 Text-to-Speech

To mitigate data scarcity, we significantly expanded our training set with synthesized audio from various TTS sources, following the strategies detailed in Section 4.3. A summary of the augmented data is provided in Table 8.

Table 7: WER results on Track 2 for with and without the Special token setting.

Setting	Step	WER
Baseline (BS) (hakka-whisper)	20000	9.31%
Special token (hakka-whisper)	30000	13.09%
Special token (openai whisper)	30000	12.82%

Table 8: Summary of synthesized TTS datasets

Data source	Nb. of Entries
External	64,950
Internal	28,046
Total	**92,996**

Ablation Study on Short-Utterance Mismatch. A key concern with synthetic data is the potential for distributional mismatch with the official dataset. We identified a significant difference in utterance length: our externally sourced, dictionary-based TTS data consists of very short clips (mean duration of 1.32 s), whereas utterances in the official training data are much longer (mean duration of 8.4 s).

To assess whether injecting a large volume of short clips would degrade model performance, we conducted a targeted ablation study. We created a data subset named **Half-Dict**, comprising approximately 38k dictionary-based utterances (totaling around 7 hours), carefully balanced between the Dapu and Zhao'an dialects.

The results, presented in Table 9, show that the inclusion of short TTS clips does not degrade performance; in fact, it provides a slight improvement in CER. We then supposed that these additional dictionary-derived utterances can expand lexical coverage, allowing the model to encounter more of the vocabulary likely to appear in the final test set.

6 Results

Based on our experiments and ablation studies, we submitted two distinct systems to the FSR-2025 challenge. The final configurations on the final dataset are summarized in Ta-

Table 9: CER result for the short-utterance (dictionary TTS) ablation.

Setting	CER (%)
Baseline	4.33
+ Half Dict.	**4.18**

Table 10: The final result on the Final dataset for the FSR-2025-challenge.

Track	Baseline	Final Result
Track 1 (CER)	10.45 %	**8.99%**
Track 2 (WER)	20.02%	**19.22%**
Track 2 (WER) (no tone value)	-	**12.36%**

ble 10.

6.1 Track 1: Recommended Characters (CER)

For Track 1, our system, which is built upon `hakka-whisper` and enhanced with our full suite of data augmentation and dialect conditioning techniques, achieved a final CER of 8.99% on the official test set. This performance secured the second-place rank among all participating teams (Fig. 2) and represents a significant 19.8% relative error reduction compared to the third-place team.

6.2 Track 2: Hakka Pinyin (WER)

For Track 2, our final system used the general-purpose `whisper-large-v3`, which outperformed the Hakka-fine-tuned variant in development. On the official test set, it achieved a WER of 19.22%, ranking third among all teams (Fig. 3). Under the competition's tone-ignored metric, the error rate further decreased to 12.36%.

7 Conclusion

This work presented a comprehensive approach for the ASR task of the under-resourced Dapu and Zhao'an dialects of Taiwanese Hakka for the FSR-2025 challenge. By integrating multiple data augmentation and robustness techniques including SpecAugment, dialect-aware special tokens, TTS augmentation, noise/reverberation mixing, and speed perturbation, our systems effectively mitigated the challenges posed by limited training

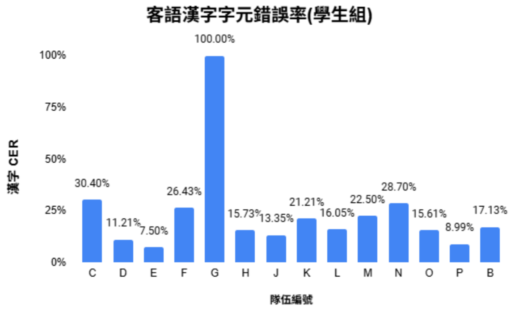


Figure 2: Official CER results for the FSR-2025 student group. Our team ranked second.

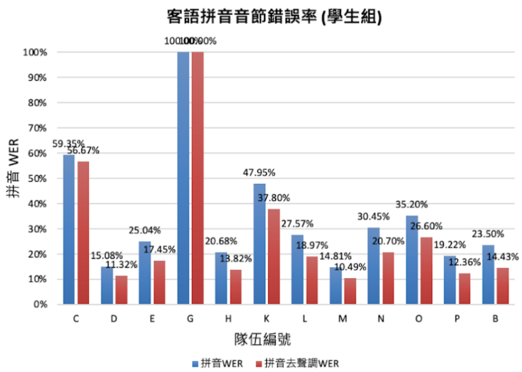


Figure 3: Official WER results for the FSR-2025 student group. Our team ranked third.

data and domain mismatch.

This report also consisted of experimental results that demonstrated substantial improvements in both CER and WER, with our Track 1 system achieving 8.99% CER (2nd place in academic groups) and our Track 2 system achieving 19.22% WER (3rd place in academic groups), further reduced to 12.36% without considering tone value. Ablation studies confirmed that each component contributed positively to overall performance.

These results highlight the effectiveness of a combined augmentation and robustness strategy for low-resource ASR, providing a practical path toward robust recognition for Hakka dialects and offering insights applicable to other under-resourced languages.

Acknowledgments

We thank the National Center for High-performance Computing (NCHC), Taiwan, for providing essential computational and storage resources.

References

Matthew Baas and Herman Kamper. 2021. Voice conversion can improve asr in very low-resource settings. *arXiv preprint arXiv:2111.02674*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

Li-Wei Chen, Hung-Shin Lee, and Chen-Chi Chang. 2024. Voxhakka: A dialectally diverse multi-speaker text-to-speech system for taiwanese hakka. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5884–5888. IEEE.

FormoSpeech. 2025. whisper-large-v3-taiwanese-hakka. `https://huggingface.co/formospeech/whisper-large-v3-taiwanese-hakka`. Accessed: 2025-09-10.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

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 transactions on audio, speech, and language processing*, 29:3451–3460.

Shao-Syuan Huang, Kuan-Po Huang, Andy T Liu, and Hung-Yi Lee. 2025. Enhancing multilingual asr for unseen languages via language embedding modeling. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Andrei Jarca, Florinel-Alin Croitoru, and Radu Tudor Ionescu. 2024. Cbm: Curriculum by masking. *arXiv preprint arXiv:2407.05193*.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. 2025. Kimi-audio technical report.

Rui Li, Guodong Ma, Dexin Zhao, Ranran Zeng, Xiaoyu Li, and Hao Huang. 2022. A policy-based approach to the specaugment method for low resource e2e asr. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 630–635. IEEE.

Hongxuan Lu and Biao Li. 2024. Sample adaptive data augmentation with progressive scheduling. *arXiv preprint arXiv:2412.00415*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Puyuan Peng and Brian Yan. 2023. Prompting the hidden talent of web-scale speech models for zero-shot task generalization.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Chih-Kai Yang, Kuan-Po Huang, and Hung-yi Lee. 2024. Do prompts really prompt? exploring the prompt understanding capability of whisper. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1–8. IEEE.

Hongli Yang, Yizhou Peng, Hao Huang, and Sheng Li. 2025. Adapting whisper for parameter-efficient code-switching speech recognition via soft prompt tuning. *arXiv preprint arXiv:2506.21576*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

8 Appendix

8.1 Strategy Selection and Cross-Track Evaluation

We initially observed consistent improvements in the Character Error Rate (CER), which suggested that joint training would not cause conflicts across different objectives. Motivated by this trend, we applied the same strategy to Track 2 and evaluated Word Error Rate (WER). In practice, however, performance on the pinyin-based task degraded, indicating that *speech-to-Chinese-character* and *speech-to-pinyin* are inherently different and should not necessarily share identical optimization recipes.

During early screening, SpecAugment was found to hurt WER and was therefore excluded from the final Track 2 configuration. As summarized in Table 11, both dialect special tokens and TTS (Half-Dict.) improved or maintained character-level recognition but further reduced WER relative to the baseline. Given competition timelines and limited compute, we could not conduct deeper; we leave these analyses to future work.

Table 11: Overview of strategies in our FSR-2025 implementation. "–" indicates the setting was excluded from Track 2 after early negative results.

Strategy	CER	WER
Baseline	4.33%	9.31%
Baseline+SpecAug	3.13%	–
Baseline+Special token	3.48%	12.82%
Baseline+TTS (Half Dict.)	4.18%	13.76%

8.2 Special tokens with additional tones

This additional part is for we have rich Hailu (304.123 hrs) and Sixian (312.369 hrs) dialect data compared to Dapu (34 hrs) and Zhaoan (34 hrs), we conducted an additional experiment with special tokens. We implemented two configurations: a baseline without dialect information and a hard-prompt approach that prepends dialect-specific tokens (e.g., `<|dialect_sixian|>`, `<|dialect_hailu|>`) to the input sequence. The intention was to leverage these high-resource dialects to improve the model's generalization to low-resource dialects and reduce overfitting. Under the CER (track 1) setting, the baseline achieved 5.00% while the hard-prompt system obtained 5.54%. Although explicit dialect prompts did not improve performance at this stage, these results provide insights for future approaches such as weighted dialect embeddings or automatic dialect inference.

A Channel-Aware Anomaly-Guided Data Augmentation Framework for the FSR-2025 Hakka Speech Recognition Challenge

Siang-Ting Lin, Arthur Hao, Chiun-Yu Hua, Kuan-Tang Huang, and Berlin Chen
National Taiwan Normal University, Taiwan
{61347114s, 41247050s, 614k0009c, 61347002s, berlin}@ntnu.edu.tw

Abstract

The Formosa Speech Recognition Challenge 2025 (FSR-2025) focuses on Taiwanese Hakka, a low-resource language with limited data diversity and channel coverage. To address this challenge, we propose a channel-aware, data-centric framework that leverages multilingual foundation models to mitigate mismatches between field recordings and training data. Our method integrates unsupervised anomaly detection and channel-conditioned augmentation to enhance data representativeness before ASR fine-tuning, aiming to explore the potential for improving robustness in low-resource Hakka speech recognition.

Keywords: Hakka Speech Recognition, Low-Resource Language, Domain Adaptation, Anomaly Detection, Data Augmentation

1 Introduction

Hakka remains a low-resource language for Automatic Speech Recognition (ASR). The challenge goes beyond limited overall data availability: it is particularly difficult to collect speech that adequately covers diverse real-world conditions, such as background noise, channel variability, and device or room effects. As a result, existing systems trained on insufficiently diverse data often lack robustness to these factors, which severely undermines practical deployment (Lu et al., 2023; Yang et al., 2023; Chen et al., 2023).

To address this gap, we adopt a data-centric pipeline that leverages multilingual resources while explicitly targeting the mismatch between field recordings and training data. Concretely, we first perform channel-aware data preprocessing and augmentation, and fine-tune Whisper (Radford et al., 2022) on the curated data.

An overview of the proposed data-centric pipeline is illustrated in Figure 1. The framework comprises three main modules corresponding to the system workflow: (1) **Target Data Selection**, where an anomaly detector based on Deep SVDD (Ruff et al., 2018) scores the test set to identify anomalous samples; (2) **Simulation Data Generation**, which employs CADA-GAN (Wang et al., 2025) to synthesize channel-aware augmented data; and (3) **ASR Fine-Tuning**, where the augmented and original training sets are jointly used to fine-tune the Whisper-based model. This three-stage pipeline unifies anomaly detection, simulation, and fine-tuning in a data-centric manner to address the channel mismatch problem in low-resource Hakka ASR.

Our design is pragmatic for the Hakka-in-the-wild setting: distribution shifts are often dominated by channel and environmental factors, e.g. device, room, reverberation, intermittent noise, which are only weakly captured by content- or speaker-centric supervision. We therefore separate two roles. First, a *channel-aware anomaly detector* operates on utterance embeddings to surface target-domain risks *without labels* and to prioritize channel conditions that the original training set undercovers. Second, a *channel-aware augmentation* stage consumes these rankings/statistics to expose the model to those under-represented conditions before fine-tuning.

Methodologically, our detector reuses an MFA-Conformer (Zhang et al., 2022) backbone to produce utterance-level embeddings, with channel supervision following prior channel-aware work. Per channel group, we adopt a lightweight two-layer Multi-Layer Per-

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 499–503
November 20-22, 2025 ©2025 Association for Computational Linguistics

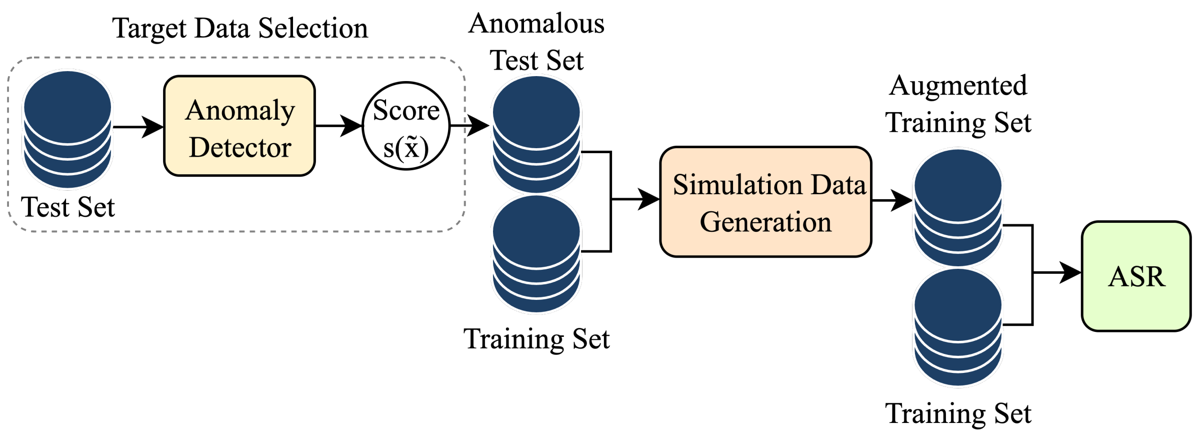


Figure 1: Overview of our data-centric pipeline. After training the anomaly detector with the training set using Deep SVDD, the detector scores the test set to filter anomalous samples (see Sec. 3.1 Target Data Selection). The selected data are then used to drive channel-aware simulation and data augmentation (see Sec. 3.2 Simulation Data Generation). Finally, the augmented and original training data are combined to fine-tune the Whisper-based ASR model (see Sec. 4.2 Model Configuration).

ceptron (MLP) with a **Soft-boundary deep SVDD** objective to score outliers; thresholds are derived from training-score quantiles and lightly calibrated on the target domain without retraining. This keeps the pipeline simple, label-free on the target side, and aligned with downstream augmentation and Whisper-based fine-tuning.

Contributions. (1) A channel-aware, data-centric pipeline for low-resource Hakka ASR that couples unsupervised detection with targeted augmentation prior to Whisper-Large fine-tuning. (2) A per-channel anomaly detector (MFA-Conformer embeddings + **Soft-boundary deep SVDD**) with bounded, no-retraining calibration to steer flag rates toward a target band. (3) An end-to-end recipe that prioritizes under-covered channel conditions and demonstrates improved robustness, evaluated with character error rate (CER) in realistic Hakka settings.

2 Background

2.1 Anomaly Detection

Anomaly detection identifies samples that deviate from the prevailing data distribution and is widely used in fraud, security, and industrial monitoring (Chandola et al., 2009; Schölkopf et al., 1999). In our low-resource Hakka automatic speech recognition (ASR) task, its role is to *surface target-domain risks without labels* and *prioritize channel conditions* that the original training set under-covers. Concretely, we use it to: (i) group utterances by channel, (ii) score and flag outliers per channel, and (iii) hand off ranked items/statistics to the downstream channel-aware augmentation stage (Sec. 3.1).

Why channel supervision (and how it relates to this task). Utterances are encoded by an MFA-Conformer backbone. In line with channel-aware work such as CADA-GAN (Wang et al., 2025), we train the encoder with *channel supervision* and at deployment reuse the learned channel discriminator to assign each utterance to a channel group. We adopt channel supervision as a pragmatic match to anticipated sources of shift in field Hakka recordings: prior channel-aware studies indicate it can emphasize channel factors and partially *disentangle* them from speaker or linguistic content. We do not claim general superiority over speaker-, phonetic-, or noise-type supervision; rather, this choice aligns with the channel-conditional analysis and augmentation used in our pipeline.

2.1.1 Deep SVDD

Deep SVDD (Ruff et al., 2018) learns an end-to-end hypersphere in a task-specific feature space so that *normal* data lie inside while violations indicate anomalies. We adopt the soft-bound variant with a lightweight two-layer MLP.

Score. Let $\mathbf{x}$ be the encoder embedding for a sample assigned to group g. For the detector we standardize $\mathbf{x}$ and apply PCA to 128 dimensions: $\tilde{\mathbf{x}} = \mathrm{PCA}_{128}(\mathrm{Standardize}(\mathbf{x}))$. With a two-layer MLP f_θ (hidden 128, output 64) and group center $\mathbf{c}_g$, the anomaly score is

$$s(\tilde{\mathbf{x}}) = \left\| f_\theta(\tilde{\mathbf{x}}) - \mathbf{c}_g \right\|_2^2. \quad (1)$$

Soft-boundary Deep SVDD loss. Within group g we optimize

$$\mathcal{L}_g(\theta, R_g) = R_g^2 + \frac{1}{\nu_g} E\big[\max\big(0,\, s(\tilde{\mathbf{x}}) - R_g^2\big)\big] + \frac{\lambda}{2} \sum_l \|W^l\|_F^2, \quad (2)$$

where R_g is the radius, $\nu_g \in (0,1)$ trades tightness vs. violations, and $\{W^l\}$ are layer weights (implemented via AdamW weight decay). After each epoch, R_g^2 is set to the $(1-\nu_g)$ quantile of *training* scores; the decision threshold is $\tau_g = R_g^2$. A test utterance is anomalous in group g iff $s(\tilde{\mathbf{x}}) > \tau_g$. For cross-group prioritization we use a rarity indicator computed against each group's training-score distribution (no label usage), which feeds the channel-aware augmentation in Sec. 3.1.

3 Method

3.1 Target Data Selection

Scope. The unlabeled target domain (final test audio) is used solely for *unsupervised* scoring, per-channel thresholding, and ranking; no labels are accessed and no model parameters are updated with target data.

Pipeline. (1) *Channel grouping.* Reuse the channel-supervised encoder (Wang et al., 2025) to assign each utterance to a group g (grouping uses original encoder embeddings). (2) *Detector features.* For Deep SVDD we standardize embeddings and apply PCA to 128 dimensions (PCA=128). (3) *Detector model.* Within each group, train a two-layer MLP f_θ (hidden 128, output 64). Let $\mathbf{x}$ be the encoder embedding and $\tilde{\mathbf{x}} = \mathrm{PCA}_{128}(\mathrm{Standardize}(\mathbf{x}))$. Define $\mathbf{z} = f_\theta(\tilde{\mathbf{x}})$ and the group center $\mathbf{c}_g$ (mean of training $\mathbf{z}$).

Thresholding and calibration (no retraining). After each epoch we set R_g^2 to the $(1-\nu_g)$ quantile of *training* scores in group g; the decision threshold is $\tau_g = R_g^2$. At test time we keep f_θ and $\mathbf{c}_g$ fixed and adjust only ν_g (hence τ_g) within bounds (e.g., $[0.01, 0.10]$) to steer the group's flag rate toward a target band ($\sim$5%). This *auto-calibration* accommodates train—test mismatch without updating model parameters.

Decision and ranking. A test utterance in group g is anomalous iff $s(\tilde{\mathbf{x}}) > \tau_g$. For cross-group prioritization we use a stable ordering: (1) anomalous first $\Rightarrow$ (2) smaller rarity indicator (tail probability, computed against the group's training scores) $\Rightarrow$ (3) larger s. The resulting per-channel flag rates and thresholds $\{(\nu_g, \tau_g)\}$ guide channel-aware augmentation to expose the ASR model to characteristics under-covered by the original training set.

3.2 Simulation data generation

We adopt CADA-GAN, a Channel-Aware Domain-Adaptive Generative Adversarial Network proposed by Wang et al. (Wang et al., 2025). The model is specifically designed to address channel mismatch in ASR by generating augmented speech data conditioned on channel characteristics. In our framework, CADA-GAN is used to synthesize additional training utterances, enriching the channel diversity of the training set.

Channel encoder: The data identified by the Deep SVDD method are used as the target source and processed by the MFA conformer to extract channel-aware representations. These representations are subsequently employed in the generator via Feature-wise Linear Modulation (FiLM)(Perez et al., 2018), where they are transformed into weights and biases to modulate the data generation process.

Generator and Discriminator: During this process, the generator integrates the encoded source data with FiLM to synthesize simulated data, while the discriminator enforces consistency between the generated data and both the intrinsic characteristics of the original source data and the embeddings of the target data.

4 Experimental Setup

4.1 Dataset

	Sentences	Hours
Train	21,879	52
Eval	5,470	8
Test(warm-up)	4,404	10
Total	31,753	70

Table 1: Dataset statistics of the FSR-2025-Hakka corpus.

We use the FSR-2025-Hakka corpus as our primary dataset. The train set contains a total of 60 hours of speech, evenly divided between two dialects: Dapu and Zhao'an (30 hours each). From this corpus, 20% of the data is randomly selected as the Eval set, while the remaining 80% is used as the Train set. The Test set consists of 10 hours of speech released for the warm-up phase, which is employed to evaluate inference performance after fine-tuning. The dataset composition is summarized in Table 1.

4.2 Model Configuration

We employed OpenAI's Whisper-Large model as our base architecture. The model configuration consisted of the following components:

Pre-trained Model We utilized the "openai/whisper-large" pre-trained model, which provides robust multilingual speech recognition capabilities. To optimize training efficiency and prevent catastrophic forgetting of learned features, we applied encoder freezing strategy, allowing only the decoder parameters to be updated during fine-tuning.

Training Strategy Our training approach employed the Seq2SeqTrainer framework. We set the batch size to 8 and accumulated gradients over 8 steps, yielding an effective batch size of 64. The model was optimized with a learning rate of 1×10^{-4}, scheduled linearly with 1,000 warmup steps. Training proceeded for 20 epochs with early stopping based on validation performance. To mitigate overfitting, we applied a weight decay of 0.01, while gradient clipping was enforced with a maximum norm of 1.0. For efficiency, we enabled mixed-precision training (FP16) and activated gradient checkpointing to reduce memory consumption.

The training dataset comprised Hakka speech data from Dapu and Zhao'an dialect variants with total 60hr data

4.3 Evaluation Metrics

Following established practices in automatic speech recognition evaluation , we employed Character Error Rate (CER) as our primary evaluation metric, which is particularly suitable for Chinese languages including Hakka.

CER The CER measures recognition accuracy at the character level and is computed as:

$$CER = \frac{S + D + I}{N} \times 100\%, \tag{3}$$

where S represents character substitutions, D represents deletions, I represents insertions, and N is the total number of characters in the reference transcript.

5 Results

Table 2 shows the CER of different training settings. Without preprocessing, the baseline system achieved a CER of 16.07%. Through systematic experimentation with different augmentation ratios, we identified 13% augmented data as the optimal configuration, yielding a CER of 15.13% when the augmented samples were generated to simulate the test set channel characteristics.

These results demonstrate that our proposed augmentation method achieves substantial performance improvement, with the optimal 13% augmentation ratio providing a 0.94 percentage point reduction in CER compared to the baseline, confirming the effectiveness of our channel simulation approach.

Method	CER
w/o Preprocessing	16.07%
Add 13% augmented data	15.13%

Table 2: CER Compare Table.

6 Conclusion and Future Work

This work presents a channel-aware, data-centric pipeline that combines unsupervised anomaly detection with targeted augmentation to address channel mismatch in low-resource Hakka ASR. By incorporating 13%

channel-simulated data, our approach reduces CER to 15.13%, achieving a 0.94-point improvement over the baseline. Our results demonstrate enhanced model robustness in realistic, noisy environments, validating the effectiveness of channel-focused augmentation.

For future work, we plan to extend the preprocessing pipeline to include semantic- and noise-specific analysis, enabling more fine-grained supervision of both linguistic and acoustic variations. In particular, long-duration noise segments, which may currently be misclassified as channel shifts, will be addressed through targeted refinement. Moreover, we will further investigate the role of FiLM modulation, as excessive influence from the generator may overpower the modulation process and reduce the contribution of source data, potentially limiting the effectiveness of synthetic augmentation.

7 Limitation

Data scale and coverage. Hakka remains low-resource; the amount and channel diversity of transcribed training audio constrain fine-tuning effectiveness. Coverage gaps (devices/rooms/reverberation patterns) limit how well Whisper-Large can adapt, even with targeted augmentation.

References

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Li-Wei Chen, Kai-Chen Cheng, and Hung-Shin Lee. 2023. The north system for Formosa speech recognition challenge 2023. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 386–389, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Hao-Chien Lu, Chung-Chun Wang, Jhen-Ke Lin, and Tien-Hong Lo. 2023. The NTNU ASR system for Formosa speech recognition challenge 2023. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 397–402, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. 1999. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Chien-Chun Wang, Li-Wei Chen, Cheng-Kang Chou, Hung-Shin Lee, Berlin Chen, and Hsin-Min Wang. 2025. Channel-aware domain-adaptive generative adversarial network for robust speech recognition.

Tzu-Ting Yang, Hsin-Wei Wang, Meng-Ting Tsai, and Berlin Chen. 2023. The NTNU super monster team (SPMT) system for the Formosa speech recognition challenge 2023 - Hakka ASR. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 414–422, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung yi Lee, and Helen Meng. 2022. Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification.

The SLAM Hakka ASR System for Formosa Speech Recognition Challenge 2025

CHIH-HSI CHEN
Department of Computer Science and Information Engineering, National Cheng Kung University
chenbrian93@gmail.com

PEI-JUN LIAO*
Institute of Information Science, Academia Sinica
newsboy3423@iis.sinica.edu.tw

CHIA-HUA WU
Institute of Information Science, Academia Sinica
maxwu@iis.sinica.edu.tw

PANG-CHENG WU
National Yang Ming Chiao Tung University
andy610125@gmail.com

HSIN-MIN WANG*
Institute of Information Science, Academia Sinica
whm@iis.sinica.edu.tw

***Corresponding author**

Abstract

In recent years, large-scale pre-trained speech models such as Whisper have been widely applied to speech recognition. While they achieve strong performance on high-resource languages such as English and Mandarin, dialects and other low-resource languages remain challenging due to limited data availability. The government-led "Formosa Speech in the Wild (FSW) project" is an important cultural preservation initiative for Hakka, a regional dialect, where the development of Hakka ASR systems represents a key technological milestone. Beyond model architecture, data processing and training strategies are also critical. In this paper, we explore data augmentation techniques for Hakka speech, including TTS and MUSAN-based approaches, and analyze different data combinations by fine-tuning the pre-trained Whisper model. We participated in the 2025 Hakka FSR ASR competition (student track) for the Dapu and Zhaoan varieties. In the pilot test, our system achieved 7th place in character recognition (CER: 15.92) and 3rd place in pinyin recognition (SER: 20.49). In the official finals, our system ranked 6 in Hanzi recognition (CER: 15.73) and 4 in Pinyin recognition (SER: 20.68). We believe that such data augmentation strategies can advance research on Hakka ASR and support the long-term preservation of Hakka culture.

Keywords: Hakka, ASR, Low Resource

1 Introduction

In recent years, Taiwan has actively invested in the preservation and development of national languages, and has promoted mother-tongue education in primary and secondary schools. In addition to Taiwanese (Southern Min), Indigenous languages, and the languages of Southeast Asian new immigrants, Hakka has also been a major focus. To encourage its daily use, teaching, and revitalization, the "Formosa Speech in the Wild (FSW) project" has launched dialect preservation initiatives, including the organization of the FSR community competition for Hakka automatic speech recognition (ASR). This shared task provides timely benchmarks and resources, with the second edition held in 2025. Hakka belongs to the Sinitic branch and encompasses multiple regional phonological systems. In particular, the Dapu and Zhaoan varieties used in the 2025 competition differ not only in segmental systems but also in prosody, such as tone and duration. Over the long term, the lack of a widely adopted writing system, combined with the declining use of Hakka among

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 504–511
November 20-22, 2025 ©2025 Association for Computational Linguistics

younger generations, has restricted the availability of annotated corpora. From a cultural perspective, however, Hakka is central to the transmission of Hakka heritage; from a technological perspective, ASR can support pronunciation assessment and computer-assisted language learning.

We approach Hakka ASR as a data-centric transfer learning challenge, emphasizing the strategic fine-tuning of the powerful general-purpose foundation model Whisper-large-v3 (Radford et al., 2023) to enhance performance on Hakka corpora. We chose Whisper as our backbone model due to its verified multilingual capability, stability in transfer learning, and feasibility on commonly available GPU hardware.

To address the limited training data for the Dapu and Zhaoan dialects, we adopted several strategies:
(i) extending the training set with synthetic speech generated by a Text-to-Speech (TTS) system;
(ii) collecting audio-text pairs from publicly available Hakka learning platforms, following the procedure described by Chen et al. (2023), to construct additional training data for the Dapu and Zhaoan dialects (restricted to Hanzi transcriptions);
(iii) incorporating speech and text from Hakka radio broadcasts in the same dialects. For data augmentation, we first applied MUSAN (Snyder et al., 2015) to inject random noise, and further employed Audiomentations (Ronny, 2020) to introduce dynamic perturbations within each training batch, thereby improving model robustness. In the 2025 FSR Hakka ASR competition (student track), our system achieved 7th place in Hanzi recognition (Character Error Rate, CER: 15.92) and 3rd place in Pinyin recognition (Syllable Error Rate, SER: 20.49) during the pilot test. In the official finals, our system ranked 6th in Hanzi recognition (CER: 15.73) and 4th in Pinyin recognition (Word Error Rate, WER: 20.68).

The following sections describe in detail our strategy for leveraging Whisper, the methods used for data augmentation and corpus expansion, and the full set of experimental results, highlighting the effectiveness and limitations of each approach. Finally, we discuss the broader implications of our findings for speech technology, especially in the context of preserving and revitalizing cultural languages. Through this study, we aim to provide methodological insights and practical tools for the future development of Hakka ASR and other low-resource language technologies.

2 Model Architecture

We use the fine-tuned Whisper model as our final submission to the competition. In addition, we perform fine-tuning on LLaMA-Omni for comparison. The details and descriptions of both models are presented below.

2.1 Whisper

In this competition, we adopt Whisper as the backbone model, following the approach of Lu et al. (2023). Whisper is an encoder-decoder ASR model pretrained on large-scale speech-text corpora. Our fine-tuning strategy focuses specifically on the decoder for the following reasons:

(i) We aim to fully leverage the pretrained knowledge on the encoder side. We assume that Whisper's encoder, which is responsible for encoding acoustic information, has strong generalization ability across different languages. Therefore, rather than fine-tuning the encoder on a small amount of target data—which might risk degrading this generalization—we retain its pretrained capacity as much as possible.

(ii) We regard the decoder as the component that adapts to the target language. From the perspective of a traditional language model, the decoder primarily handles the mapping from acoustic features to linguistic representations. Since this process must reflect the characteristics of the target language (e.g., Hanzi or Pinyin language models), fine-tuning the decoder is a reasonable and effective choice.

2.2 LLaMA-Omni

With the recent rise of large audio–language models (LALMs), such as those proposed by Zhang et al. (2023) and Chu et al. (2024), we conducted additional experiments using the

Source	Dataset Name	Usage	Duration(hr)	Description
Organizer provided	FSR-2025-Train	Train set	62.0	Official training corpus released by competition organizers.
	FSR-2025-Record	Train set	7.2	Pilot-test subset
		Test set	0.8	
	FSR-2025-Media	Train set	1.6	Pilot-test subset
		Eval set	0.2	
		Test set	0.2	
Web collected	Hakka Radio	Train set	11.0	Transcribed broadcast speech
	Hakka E-Learning	Train set	16.0	Educational reading material
TTS generated	FSR-Website-TTS	Train set	335.0	Synthetic speech from VITS trained on FSR-2025-Train.
	FSR-Media-TTS	Train set	8.0	Synthetic speech from VITS trained on FSR-2025-Train
Competition test sets	P-test	Test set	1.0	1 hr subset (0.8 Record + 0.2 Media)
	F-test	Test set	10.0	Final competition set.

Table 1 Summary of all speech datasets used in this study

LLaMA-Omni model (Fang et al., 2024). The original architecture employs an 8B large language model; however, in our implementation, we replace it with a smaller 1B-parameter LLaMA (Dubey et al., 2025) variant to better accommodate limited GPU resources. The architecture integrates the Whisper-large-v3 encoder, and a linear adapter is inserted between the encoder and the LLM to align their feature dimensions by projecting the encoder output into the LLM's embedding space.

For fine-tuning, we follow a similar strategy by freezing the Whisper encoder to preserve its pretrained capacity for extracting meaningful speech representations. The adapter and LLM components are then trained jointly, enabling the model to adapt to the downstream task. This setup allows us to leverage the robust acoustic representations from the frozen encoder while focusing computational resources on adapting the modality-bridging adapter and the large language model to the target language domain. This configuration serves as a comparative baseline against our Whisper-only fine-tuning approach.

3 Data Sources

We first remove silence segments from all speech data to avoid adverse effects on model training. In addition, all corpora—including both the organizer-provided data and our self-collected resources—are resampled to 16 kHz to ensure consistency with the model requirements. Below, we describe our data augmentation and processing methods, as well as the training mechanism for data utilization. The overall pipeline of our data processing and model fine-tuning framework is illustrated in Figure 1. A detailed summary of all datasets used in this work is provided in Table 1.

3.1 FSR Hakka Challenge

As summarized in Table 1, the datasets used in this study can be grouped into three categories:(i) official FSR corpora released by the organizers, (ii) web collected resources, and (iii) TTS-generated synthetic speech. These corpora collectively provide complementary coverage of read and spontaneous Hakka speech, forming the basis for the experiments in Section 4.

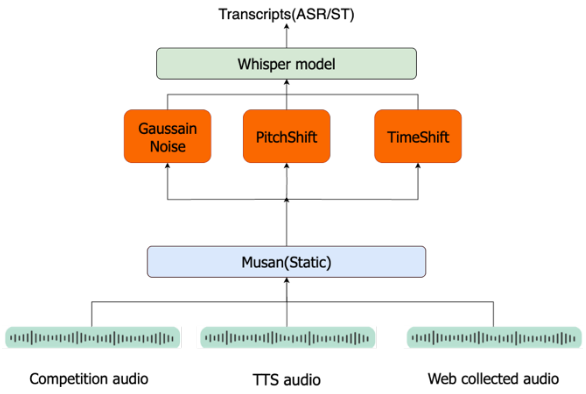


Figure 1: Overall architecture of the AS-SLAM system for Hakka ASR.

3.2 Web-Collected

We extracted speech-text pairs for both the Dapu and Zhaoan dialects from two publicly available online resources: Hakka E-learning and Hakka Radio.

The duration and usage of the web-collected datasets are summarized in **Table 1.** Specifically, The Hakka Radio corpus comprises approximately 9 hours of Dapu speech and 2 hours of Zhaoan speech, collected from broadcast programs in which native speakers discuss daily topics in a spontaneous conversational style.

These recordings exhibit diverse acoustic environments, speaker variations, and natural prosodic patterns. Owing to their broadcast nature, we hypothesize that the distribution of Hakka Radio more closely resembles that of the FSR-2025-Media subset. In contrast, the Hakka E-Learning corpus includes 8 hours each for Dapu and Zhaoan, originating from educational materials on the official Hakka E-learning platform. The utterances are primarily short, card-style sentences in which speakers read isolated words or short phrases aloud. Because of its clear articulation and relatively clean recording conditions, their corpus shares greater similarity with the FSR-2025-Record subset. Since the transcripts of both datasets are in Hanzi, they are used exclusively for the Hanzi track of the competition.

3.3 TTS-Generated

We adopt the Variational Inference Text-to-Speech (VITS) model (Kim et al., 2021) for speech synthesis, following the approach of Chen et al. (2023). During model training, we use both the official training data provided by the organizers and our self-collected resources. Separate TTS models are trained for the Dapu and Zhaoan dialects. For text prompts, we collect 150,000 example sentences from Hakka dictionary online published by the Hakka Affairs Council[1]. As summarized in Table 1, the official *FSR-2025-Train* corpus includes 123 speakers across both dialects and genders. From this pool, five speakers are randomly selected, and each generates 67 hours of speech, resulting in a total of 335 hours of synthetic data (denoted as FSR-Website-TTS). In addition, to enhance the coverage of media-style speech, we reuse the previously trained TTS models to perform speech synthesis using the transcriptions from the FSR-2025-Media dataset (1.6 hours of text). For each dialect, five speakers are randomly sampled, and each generates 1.6 hours of speech, resulting in a total of 8 hours of synthetic data (denoted as FSR-Media-TTS). Due to time constraints, only Hanzi transcriptions were used for speech synthesis.

4 Data Augmentation

We divide our data augmentation into two strategies—static and dynamic—following the two-stage approach proposed by Bhat et al. (2025), which are described as follows.

The overall workflow of both augmentation stages and their integration with the Whisper fine-tuning pipeline is illustrated in Figure 1. As shown in the figure, all audio sources—including competition data, TTS-generated data, and web-collected corpora—first pass through a static augmentation stage (MUSAN), followed by dynamic augmentations applied online during model training.

These two levels of augmentation jointly enhanced the model's robustness to noise, channel, variation, and acoustic mismatch across domains.

[1] https://hakkadict.moe.edu.tw/resource_download/

4.1 Static Data Augmentation

We employ the MUSAN (Snyder et al., 2015) toolkit, MetricAug (Wu et al., 2023), and the method proposed by Ko et al. (2023) for data augmentation, adding noise to clean speech before the training stage. The noise level is controlled by a randomly sampled signal-to-noise ratio (SNR) between 0 and 15 dB, following the configuration described in Pligin-SE (Chen et al., 2024). This static stage serves as the offline augmentation block shown in Figure 1 , ensuring that each input waveform exhibits realistic acoustic diversity prior to entering the dynamic augmentation pipeline.

4.2 Dynamic Data Augmentation

We apply the Audiomentations [2] toolkit for dynamic data augmentation. Unlike static data augmentation, this method is integrated directly into the training process. Before each sample is fed into the model, the following transformations are independently applied, following the configuration described in Dynamic Mixing (Choi et al., 2022) and Aligned Data Augmentation (Lam et al., 2021): GaussianNoise (minimum amplitude = 0.001, maximum amplitude = 0.015, probability = 0.3), TimeStretch (minimum rate = 0.9, maximum rate = 1.1, probability = 0.3), and PitchShift (minimum semitone = –2, maximum semitone = 2, probability = 0.3). This dynamic augmentation introduces greater variability during training, thereby improving the model's robustness.

5 Experimental Setup

5.1 FSR Challenge Setting

After the pilot test (stage 1 of the competition), our submitted model showed notably weaker performance on the FSR-2025-Media subset, suggesting that the model was less robust to the media distribution. To address this issue, we extended the training data by adding 1.6 hours of FSR-2025-Media and 7.2 hours of FSR-2025-Record to the original 40 hours of FSR-2025-Train. This new configuration is referred to as FSR-2025-Train-Plus, and served as the baseline for our final experiments. Building on this setup, we designed a series of extended experiments to examine the impact of additional data sources and augmentation strategies. Specifically, we trained three systems before the final submission deadline:

1. **FSR-2025-Train:** The official 40-hour training set only.
2. **FSR-2025-Train-Plus**: FSR-2025-Train+FSR-2025-Record+FSR-2025-Media with the hour combination described above.
3. **FSR**-2025-Train-Final: An extended configuration that further incorporates web-collected corpora and synthetic TTS speech.

All systems were trained with both static and dynamic data augmentation. The remaining experimental variants and comparative results are presented in Section 4.5 (Ablation Study).

In the Pinyin track, we did not incorporate self-collected corpora or TTS-generated data; instead, the system relied solely on the organizer-provided datasets. During the pilot test, training was conducted exclusively on the organizer-provided data with both static and dynamic augmentation strategies applied. In this setting, 20% of the 40-hour dataset was held out as the validation set, resulting in 32 hours of original speech data used for training. In the final stage, due to time constraints, we combined the 40-hour FSR-2025-Train dataset with the full 8 hours of FSR-2025-Record and the complete 2-hour FSR-2025-Media dataset for final model training, without a separate validation set; the model from the last training checkpoint was directly used for prediction. In both tracks, our final submission model was based on the Whisper architecture, fine-tuned at the decoder.

5.2 Model Training Details

For all competition submissions, we fine-tuned the Whisper-large-v3 model for 10 epochs, with a learning rate of 1e-5 following Whisper-LM (Zuazo et al., 2025) and an accumulated batch size of 64. For comparison, the LLaMA-Omni model was trained with a learning rate of 1e-4, an accumulated batch size of 12, and for 10 epochs

[2] https://github.com/iver56/audiomentations

in total. All experiments were conducted on Ubuntu using an NVIDIA RTX 3090 GPU.

6 Experiment Results

6.1 Pilot Test Results

In the pilot test stage, our system achieved a CER of 15.92% on the Hanzi track (Figure 2) and an SER of 20.49% on the Pinyin track (Figure 3) in the official student division results, ranking 7th and 3rd, respectively. The pilot test dataset consists of two subsets: FSR-2025-Record and FSR-2025-Media. To expand the preliminary data for the final competition, we constructed a new test set by combining 0.2 hours from FSR-2025-Media and 0.8 hours from FSR-2025-Record, referred to as P-Test. We then evaluated both the Whisper and LLaMA-Omni models, trained on the final competition training data, using this 1-hour Hanzi test set. The results are presented in Table 2. As shown, under limited training data conditions, Whisper still clearly outperforms LLaMA-Omni. We attribute this to the fact that automatic speech recognition (ASR) is Whisper's original pretraining objective, whereas LLaMA-Omni is designed for more general multimodal purposes. Consequently, Whisper holds a stronger advantage in ASR-specific tasks.

6.2 Final Competition Results

In the finals, our system achieved a CER of 15.73% on the Hanzi track and, for the Pinyin track, a WER of 20.68% and a tone-removed WER (WER^) of 13.82% in the official student division results, ranking 6th and 4th, respectively. The rankings and corresponding error rates are summarized in Table 3.

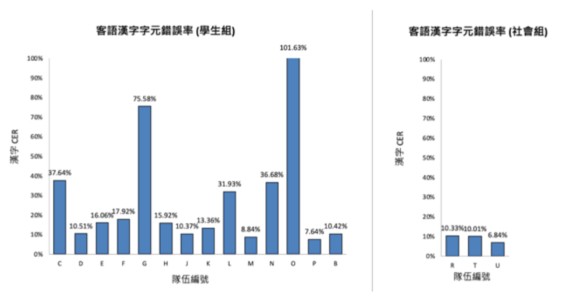


Figure 2: CER results and rankings on the Hanzi track in the pilot test. Our team, labeled as "H," participated in the student division.

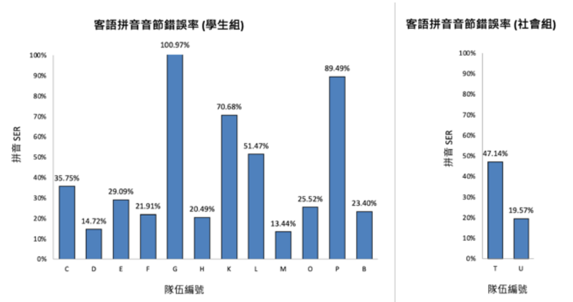


Figure 3: SER results and rankings on the Pinyin track in the pilot test. Our team, labeled as "H," participated in the student division.

Model	CER
LLaMA-Omni	4.24
Whisper	2.54

Table 2: CER Results on the Hanzi Track of the P-Test for LLaMA-Omni and Whisper. Both models were trained under the same data configuration as FSR-2025-Train-Final

6.3 Ablation Study

We conducted a series of ablation experiments to examine the contribution of each data source, as summarized in **Table 4**. We observed that in the P-Test, simply adding the FSR-Website-TTS data led to a performance decline, whereas adding the FSR-Media-TTS data resulted in improved performance on the P-Test but showed the opposite trend on the F-Test. All configurations employed both static (MUSAN) and dynamic (Audiomentations) data augmentation, and were evaluated on the 1-hour **P-Test** and the **F-Test**.

Moreover, combining both types of TTS data did not yield any complementary effect on either test. We speculate that this discrepancy may be attributed to the distributional differences between the two types of TTS data. In contrast, both Hakka Radio and Hakka E-Learning contributed significant improvements on the P-Test and F-Test, with even greater gains when the two were combined.

Notably, on the F-Test, using only these two datasets outperformed all other data combinations. In the Web-collected data experiments, we further observed that adding Hakka Radio yielded better performance than adding Hakka E-Learning. Interestingly, incorporating only Hakka E-Learning caused a performance drop on the P-Test but showed improvement on the F-Test.

When both Hakka E-Learning and Hakka Radio were added together, the performance on the P-Test was slightly worse than using Hakka Radio alone, whereas on the F-Test, the two datasets exhibited complementary effects. We speculate that this is because the data d istribution of Hakka E-Learning differs considerably from that of general media data, while Hakka Radio demonstrates higher generalizability. This effect may also be influenced by the higher proportion of media data in the P-Test compared with the F-Test.

Hanzi		Pinyin		
Rank	CER	Rank	WER	WER^
6	15.73%	4	20.68%	13.82%

Table 3: In the final results of the Hanzi and Pinyin tracks, the evaluation metric for the Hanzi track is CER, while that for the Pinyin track is WER. WER^ denotes the WER evaluated after tone removal.

	CER	
	P-Test	F-Test
FSR-2025-Train	20.10%	27.79%
FSR-2025-Train-Plus	3.40%	17.54%
+ FSR-Website TTS (1)	3.78%	17.07%
+ FSR-Media TTS (2)	3.29%	17.33%
+ (1) + (2)	3.5%	17.50%
+ Hakka Radio (3)	3.07%	15.04%
+ Hakka E-Learning (4)	3.72%	16.71%
+ (3) + (4)	2.55%	14.58%
+ (1) + (2) + (3) + (4)	2.54%	15.73%

Table 4: Comparison of training solely on the original FSR 2025 dataset versus augmenting it with TTS, Hakka E-Learning, and Hakka Radio, evaluated on the Hanzi track of 1-hour Pilot test in the pilot test (P-Test) and Final test set (F-Test) in terms of CER.

	WER
	F-Test
FSR-2025-Train	32.60%
FSR-2025-All	20.68%

Table 5: Comparison of the results on the Final Pinyin Track using the training data from the preliminary round (FSR-2025-Train) and the training data used for the Final Pinyin Track (FSR-2025-All).

For the Pinyin track, we used all FSR-2025-Media and FSR-2025-Record data in the final stage, while keeping the remaining configurations identical to FSR-2025-Train-Plus. We refer to this training set as FSR-2025-All. The models trained with both the pilot test training data and this final training set were evaluated on the final test set, and the WER results are shown in Table 5. We observed a significant improvement after adding the additional training data, suggesting that future research could further enhance model performance by expanding the amount of Pinyin training data.

7 Conclusion

In this competition, we investigated the use of various Hakka datasets and conducted preliminary experiments with existing ASR models such as Whisper and LLaMA-Omni. Our results provide initial evidence that Whisper may outperform LLaMA-Omni for ASR tasks in low-resource languages. In the finals, we achieved 6th place in the Hanzi track and 4th place in the Pinyin track. Moving forward, we plan to explore integrating data across different dialects and experimenting with more model combinations, with the goal of making further progress in low-resource language research.

Acknowledgments

We would like to express our sincere gratitude to Dr. Hung-Shin Lee from North Co., Ltd., Taiwan, for providing the web-collected data.

References

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning* (pp. 28492-28518). PMLR.

L.-W. Chen, K.-C. Cheng, and H.-S Lee. 2023. The North System for Formosa Speech Recognition Challenge 2023. *In Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023).*

H.-C. Lu, C.-C. Wang, J.-K. Lin, and T.-H. Lo 2023. The NTNU ASR System for Formosa Speech Recognition Challenge 2023. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*.

D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu. 2023. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In *Findings of the Association for Computational Linguistics: ACL 2023*, (pp. 1469-1483).

Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, and others. 2024. Qwen2-Audio Technical Report. *arXiv preprint* arXiv:2407.10759.

Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng. 2024. Llama-Omni: Seamless Speech Interaction with Large Language Models. *In Proceeding of the 13th International Conference on Learning Representations (ICLR 2025).*

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, and R. Ganapathy 2024. The Llama 3 Herd of Models. *arXiv preprint* arXiv:2407.12345.

J. Kim, J. Kong, and J. Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *Proceedings of the 38th International Conference on Machine Learning (ICML).*

P.-K. Chen, B.-J. Huang, C.-T. Chen, H.-M. Wang, and J.-C. Wang. 2023. Enhancing Automatic Speech Recognition Performance Through Multi-Speaker Text-to-Speech. *In Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023).*

C. Bhat, and H. Strik. 2025. Two-Stage Data Augmentation for Improved ASR Performance for Dysarthric Speech. *Computers in Biology and Medicine*, 189, 109954.

D. Snyder, G. Chen, and D. Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. *arXiv preprint arXiv:1510.08484.*

Y.-T. Wu, and C.-C. Lee. 2023. MetricAug: A Distortion Metric-Lead Augmentation Strategy for Training Noise-Robust Speech Emotion

Recognizer. *In Proceedings of INTERSPEECH 2023. International Speech Communication Association (ISCA).*

T. Ko, V. Peddinti, D. Povey, and S. Khudanpur. 2017. A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition. *In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Y. Chen, Z. Cui, Y. Gao, J. Feng, C. Deng, and S. Zhang. 2024. Plugin Speech Enhancement: A Universal Speech Enhancement Framework Inspired by Dynamic Neural Network. *arXiv preprint* arXiv:2402.12746.

S. Choi, Y. Lee, J. Park, H. Y. Kim, B.-Y. Kim, Z.-Q. Wang, and S. Watanabe. 2022. An Empirical Study of Training Mixture Generation Strategies on Speech Separation: Dynamic Mixing and Augmentation. I*n Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).*

T. K. Lam, M. Ohta, S. Schamoni, and S. Riezler. 2021. On-the-Fly Aligned Data Augmentation for Sequence-to-Sequence ASR. *Proceedings of Interspeech 2021*, (pp. 1299-1303).

X. de Zuazo, E. Navas, I. Saratxaga, and I. Hernáez Rioja. 2025. Whisper-LM: Improving ASR Models with Language Models for Low-Resource Languages. *arXiv preprint* arXiv:2503.23542.

多語預訓練模型 Whisper 在低資源語言中的挑戰與限制探討：以客語語音辨識為例

Challenges and Limitations of the Multilingual Pre-trained Model Whisper on Low-Resource Languages: A Case Study of Hakka Speech Recognition

Pei-Chi, Lan
Department of Japanese Language and Culture
Soochow University
Taiwan
lizza63159@gmail.com

Hsin-Tien, Chiang
Department of Data Science
Soochow University
Taiwan
nataliechiang92@gmail.com

Ting-Chun, Lin
Department of Data Science
Soochow University
Taiwan
Janelin454@gmail.com

Ming-Hsiang, Su
Department of Data Science
Soochow University
Taiwan
huntfox.su@gmail.com

摘要

本研究以客語語音辨識競賽為案例，探討多語預訓練模型 Whisper 在低資源語言環境下的實務表現與限制。於熱身賽階段，研究團隊（G 組）的官方成績為漢字字元錯誤率（Character Error Rate, CER）75.58%，拼音錯誤率（Syllable Error Rate, SER）100.97%；而在決賽階段，CER 與拼音錯誤率（Word Error Rate, WER）皆達 100%。透過對系統設計與執行流程的回顧分析，我們歸納出三項主要問題來源：(1) 長語音處理策略失當，僅保留首段進行解碼導致內容截斷；(2) 解碼階段的語言提示固定為「中文」，與客語辨識目標不符；(3) 資料對齊與提交檔生成流程缺乏系統化檢核，且評估設定不當。根據這些觀察，我們提出可重複應用的實務準則，涵蓋長語音處理、語言設定一致性確認與資料提交流程檢查等面向。研究結果顯示，在低資源語言的語音辨識任務中，若資料品質與流程設計未妥善控管，即使使用先進的多語預訓練模型，其效能亦可能嚴重退化。本研究強調資料與流程管理在系統開發中的關鍵角色，並為後續改進與結果重現提供具體參考。

Abstract

This study investigates the practical performance and limitations of the multilingual pre-trained model Whisper in low-resource language settings, using a Hakka speech recognition challenge as a case study. In the preliminary phase, our team (Group G) achieved official scores of 75.58% in Character Error Rate (CER) and 100.97% in Syllable Error Rate (SER). However, in the final phase, both CER and Word Error Rate (WER) reached 100%.

Through a retrospective analysis of system design and implementation, we identified three major sources of failure: (1) improper handling of long utterances, where only the first segment was decoded, causing content truncation; (2) inconsistent language prompting, fixed to “Chinese” instead of the Hakka target; and (3) lack of systematic verification in data alignment and submission generation, combined with inadequate evaluation setup.Based on these findings, we propose a set of practical guidelines covering long-utterance processing, language consistency checking, and data submission validation. The results highlight that in low-resource speech recognition tasks, poor data quality or flawed workflow design can cause severe degradation of model performance. This study underscores the importance of robust data and process management in ASR system development and provides concrete insights for future improvements and reproducibility.

關鍵字：客語、語音辨識、低資源語言、長語音處理、語言提示、資料對齊、評估指標

Keywords: Hakka, speech recognition, low-resource language, long-audio processing, language prompting, data alignment, evaluation metrics

1 研究背景

近年來，隨著多語自監督學習模型（如 Whisper、XLS-R、wav2vec 2.0）陸續問世，語音辨識技術已逐漸從高資源語言（例如英語、中文普通話）擴展至低資源語言與方言。然而，這類模型的效能仍高度依賴語料的品質與規模，以及

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 512–517
November 20-22, 2025 ©2025 Association for Computational Linguistics

標註方式與模型設定的一致性。當資料稀疏、腔調多樣或語音特徵差異顯著時，即使採用強大的預訓練模型，也可能因流程細節錯誤而導致辨識失準。

臺灣客語屬於典型的低資源語言，其內部分化為多個腔調（如四縣、海陸、大埔、詔安等），在聲學與詞彙層面皆具有顯著差異。雖然政府與學術界已推動語料蒐集與文字化工程，但可直接用於自動語音辨識（ASR）訓練的開放語料仍相對有限。因此，客語 ASR 的開發不僅需面對低資源問題，亦須同時處理多腔調資料整合、語音長度差異與標註一致性等挑戰。

本研究以 **ROCLING 2025 客語語音辨識競賽**為案例，嘗試在有限的開源語料與模型條件下建構可運作的客語 ASR 系統。研究團隊於熱身賽中使用 Whisper 模型，雖能產生一定可辨識的輸出，但在決賽階段，系統表現卻不盡理想，官方評分之 **字元錯誤率（CER）與拼音錯誤率（WER）皆達 100%**。此結果提供了一個重要的反思契機——模型性能並非僅受限於資料量或模型規模，而更容易受到流程一致性、語料對齊、語言設定與評估機制等實務因素的深刻影響。

本文旨在以此失敗案例為出發點，檢視低資源語言 ASR 系統在實作層面的潛在風險，並透過事後分析歸納出可重複應用的檢核準則，期能為後續客語及其他低資源語言的研究提供經驗基礎與改進參考。

2 語料與資料集

本研究所使用之語料與任務皆源自 ROCLING 2025 客語語音辨識競賽。競賽主辦單位提供經整理的官方客語語音資料集，內容涵蓋多位說話者、不同腔調及多樣錄音環境。所有語料均由客家委員會「臺灣客語語音資料庫」授權使用，取樣率為 16 kHz、單聲道、16-bit PCM 編碼，並附有對應之轉寫文字（客語漢字及拼音）。研究團隊僅使用主辦方提供的資料，未額外引入外部語料或語言模型，以確保與官方評測條件一致。

2.1 熱身賽資料集

熱身階段所使用的語料為 FSR-2025-Hakka-evaluation，內容包含錄製語料與媒體語料兩部分，總時長約 10 小時。語料來源為客家委員會之「臺灣客語語音資料庫」，涵蓋大埔腔與詔安腔兩種腔別，並分為男、女聲語者共 21 人。

錄製語料共計 3,458 句、約 44,744 字元、8.0 小時；媒體語料則包含大埔腔與詔安腔各約 1 小時，共 946 句、約 26,883 字元。錄音環境與設備多樣，音檔中保留部分雜訊與腔調差異，以模擬真實語音使用情境並提升模型的泛化能力。

2.2 決賽資料集

決賽階段所使用的語料為 FSR-2025-Hakka-final-release，同樣出自客家委員會「臺灣客語語音資料庫」。該語料總時長約 10 小時，包含 4,563 句語音、約 91,642 字元。音檔以亂數命名之 WAV 檔形式提供，格式為單聲道、16-bit PCM、16 kHz 取樣率。

語料涵蓋多位說話者與兩種腔調（大埔腔、詔安腔），錄音使用多種麥克風與環境條件，部分音檔長度達 20 秒。由客語教師監聽審核，僅於讀錯字時修正，保留自然發音與環境音以維持真實語音特徵。

2.3 任務定義與限制條件

本競賽的核心任務為：給定一段客語語音，系統需自動產生對應之「客語漢字」與「客語拼音」轉寫結果，並盡可能降低錯誤率。

主辦單位未提供額外語言模型或外部語料，系統須完全依賴官方訓練資料中的語音與文字對應關係進行學習。此設定可用於評估多語預訓練模型（如 Whisper）在低資源語言環境下的實際辨識能力與資料依賴程度。

3 方法

本研究以開源多語自監督模型 Whisper 為基礎，探討其在客語語音辨識任務中的實作流程與性能表現。整體方法包含五個部分：(1) 模型架構、(2) 資料前處理、(3) 訓練設定、(4) 語言提示與長語音處理策略、(5) 評估指標與分析方法。以下將逐一介紹。

3.1 模型架構

Whisper 模型為由 OpenAI 發表之 Encoder–Decoder 結構，預先以多語音資料（超過 680,000 小時）進行訓練，能同時執行語音辨識與翻譯任務。本研究採用其開源權重進行微調。

熱身階段使用 whisper-tiny 模型，以驗證可行性；決賽階段則改用 whisper-medium，期望獲得更佳的聲學表徵能力。

3.2 資料前處理

語音樣本經 WhisperFeatureExtractor 處理後，轉換為 16 kHz 的對數梅爾頻譜（log-Mel spectrogram）。標註部分取自主辦單位提供之「客語漢字」欄位，並以 WhisperTokenizer 進行編碼。為維持多腔調資料的一致性，未額外進行拼音正規化處理。訓練資料經由 Dataset.map() 生成語音與標註特徵對應，並於訓練前隨機混合不同語者，以降低說話者差異造成的偏差。

3.3 訓練設定

模型以交叉熵損失函數（cross-entropy loss）進行序列到序列學習。熱身階段的訓練設定為：max_steps = 500、learning_rate = 1e-5、batch_size = 4，未啟用驗證集（evaluation_strategy = "no"）。決賽階段因硬體資源受限且資料量增加，設定調整為 max_steps = 300、learning_rate = 5e-6、batch_size = 1，並啟用 eval_steps = 200 以保存最佳權重。

3.4 語言提示與長語音處理

在微調與推論過程中，**WhisperProcessor** 皆設定 language = "chinese"、task = "transcribe"。此策略在多語模型中常見，但於本任務中造成語言提示與實際語音不符，使客語語音被模型誤判為中文。此外，為避免記憶體溢出，決賽實作僅保留音檔的首段（約 30 秒以內）進行推論，未採用重疊解碼或片段合併。此設計雖可降低運算負擔，卻導致長語音內容被截斷，成為模型性能嚴重退化的主要原因之一。

3.5 評估指標

官方評分以漢字字元錯誤率（Character Error Rate, CER）與拼音錯誤率（Syllable Error Rate, SER）為主要指標；內部分析則另採用字元層級 CER 與字詞層級 WER 進行比較。由於客語標註未進行分詞，WER 在此情境下容易高估實際錯誤率，因此本研究以 CER 作為主要評估指標，以反映系統在不同階段的整體辨識穩定性。

4 結果與分析

本節呈現熱身賽與決賽兩階段之實驗結果，並探討模型退化的原因。

4.1 熱身賽結果

在多腔調（大埔腔與詔安腔）資料上，模型能基本完成語音轉寫。官方評分顯示，G 組在學生組中取得漢字 CER = 75.58%、拼音 SER = 100.97%（請見圖一，G 組部分）。雖然音節層級錯誤率偏高，但部分輸出仍能保留與原句相近的字形或聲韻組合，顯示模型已學得一定程度的聲學對應關係。

主要誤差來源為腔調混用與標註不一致。若以字元層級進行分析，錯誤多集中於同音或近音字的替換，顯示模型在聲學辨識上具有限度的區辨能力，但在語言層級仍受資料品質影響。

4.2 決賽結果

進入決賽後，系統性能明顯退化。官方面板顯示，G 組之 CER 與 WER 均為 100%（請見圖二，G 組部分）。分析顯示，此極端結果並非模型「學壞」，而是由流程錯誤所導致的系統性失效。

主要問題可歸納為四點：

(1) 長語音截斷問題：程式僅保留音檔首段進行推論，導致句尾與後段內容完全遺失。

(2) 語言提示不一致：解碼端固定設定 language = "chinese"，使模型傾向輸出中文或雜訊字元。

(3) 資料對齊與提交流程缺乏驗證：決賽資料以資料夾掃描方式生成 CSV，可能造成音檔與標註對應錯位。

(4) 評估指標選擇不當：未分詞的 WER 在中文及客語語境下容易誇大插入與刪除錯誤，造成評估偏差。

4.3 討論

綜合以上觀察，熱身與決賽之間的落差顯示，模型效能對資料流程極為敏感。當語言提示與標註不一致、長語音處理不當或資料對齊錯位時，系統可能完全失效。

此結果說明，低資源語言 ASR 的瓶頸不僅在於資料稀缺，更在於流程設計的正確性與一致性。若能在實務層面建立自動化檢核機制，例如對齊驗證、語言一致性檢查與重疊解碼策略，應能顯著提升低資源語言系統的穩定性與可重現性。

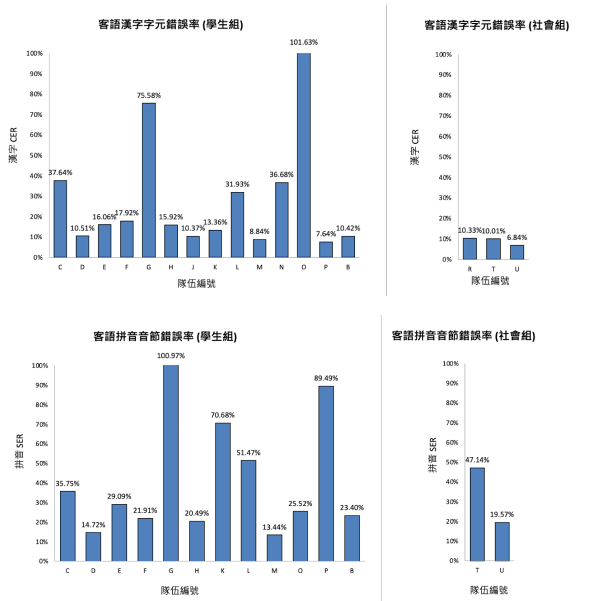


圖 1　熱身賽結果

資料來源：ROCLING 2025 客語語音辨識競賽官方網站 <https://sites.google.com/nycu.edu.tw/fsw/home/challenge-2025?authuser=0>

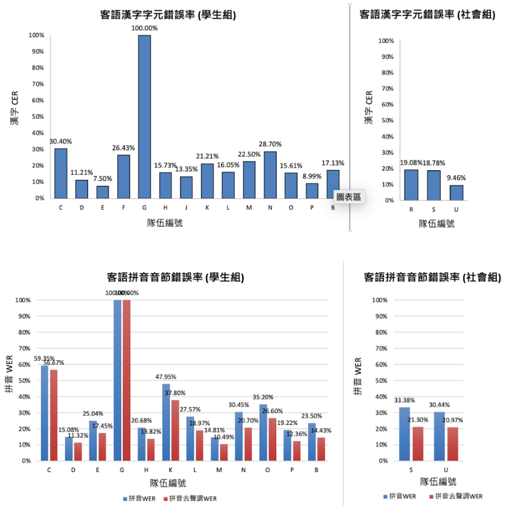


圖 2　決賽結果

資料來源：ROCLING 2025 客語語音辨識競賽官方網站 <https://sites.google.com/nycu.edu.tw/fsw/home/challenge-2025?authuser=0>

5　錯誤分析與討論

5.1　系統性錯誤來源

經追蹤程式版本、訓練記錄與提交流程後，本研究歸納出三項主要失效原因。

(1) 長語音截斷與內容遺失

決賽版本的推論程式為了節省 GPU 記憶體，僅保留每段音檔的首段進行推論。由於決賽語料中平均句長遠高於熱身賽（部分超過 15 秒），此做法直接導致語音後半段未被解碼。系統因此在句尾完全無輸出，導致 CER 與 SER 均接近 100%，屬於結構性錯誤（systemic failure）。

(2) 語料誤用與語言提示不一致

決賽階段的訓練與推論誤將熱身賽語料（大埔、詔安混合）用於模型更新，而正式決賽資料集為獨立語料，兩者語句與錄音來源完全不重疊。此一語料錯置導致模型無法對應測試集語音內容，即使運行正常也不可能產生正確轉寫。此外，Whisper 模型解碼端語言提示（language token）固定為 chinese，與客語實際語音不符，造成生成結果出現大量非預期語言。在熱身階段的輸出中，即可觀察到模型同時產生英語、日語甚至泰文拼音符號，顯示其在語言識別與字集選擇上受到提示錯置的嚴重干擾。此現象突顯多語預訓練模型在低資源語言中的脆弱性：當語言提示與聲學特徵不符時，模型傾向「回退」至高資源語言的字表與音系。

(3) 資料對齊與提交流程的不可驗證性

與熱身賽使用固定 CSV 對應不同，決賽程式改以資料夾掃描方式自動生成提交清單。若檔名排序、語者 ID 或清單順序不一致，音檔與文字對應即可能錯位。由於未設置提交前的比對驗證（例如雜湊或樣本抽查），部分輸出結果可能與官方標準答案對不上，進一步放大系統性錯誤。

5.2　評估指標與低資源語言特性

本競賽的評分方式為：熱身賽採用 CER 與 SER，決賽則以 CER 與 WER 作為主要指標。以下結果與討論均以主辦單位公布之成績為準。由於語料未經分詞處理，WER 與 SER 在計算插入與刪除時容易高估實際錯誤，因此本研究於比較趨勢時亦同時參考 CER，以獲得較穩定之評估結果。

值得注意的是，決賽階段的 CER 與 WER 雖皆達 100%，其原因並非模型退化或學習失效，而是源於語料誤用與流程錯置所造成之對齊錯誤。此結果顯示，在低資源語言任務中，資料流程設計與評估一致性對系統效能具高度敏感性與決定性影響。

5.3　低資源語言 ASR 的實務啟示

本研究的失敗案例揭示了低資源語言 ASR 在實務層面的三項主要風險。

首先是**流程治理**。資料輸入、對齊、訓練與評估各階段皆需設立驗證機制，例如版本控制、檔案雜湊、樣本抽查與提交前比對，以避免隱性錯配或資料汙染。

其次為**語言提示一致性**。多語模型在低資源語言上的正確性高度依賴提示設定；若提示語言與輸入語音不符，模型傾向回退至其訓練頻率較高的語言（如中文或英語），導致不可預期的混語輸出。

最後是**語料範圍與重疊檢核**。在資料拆分與競賽實驗中，須確認訓練集與測試集之間無重疊或錯配；於低資源語言環境中，即便少量語料誤用，也可能造成整體模型失效。

綜上所述，這些檢核與控制策略遠比單純增加資料量更能有效提升系統的穩定性與實驗可重現性。

5.4　反思

本研究顯示，「模型能力不足」並非低資源語言語音辨識系統失效的唯一原因。在此類任務中，資料準備與流程一致性往往對最終結果具有更關鍵的影響力。

本次競賽中，模型性能的崩潰並非源於技術退步，而是由實驗設計與語料治理的疏漏所導致。此案例突顯，低資源語言研究除了演算法創新之外，更需重視資料品質與流程管理。

未來的客語 ASR 研究應將語料完整性驗證與流程可追溯性納入標準研究流程，並建立跨團隊共用的檢核框架，使低資源語言研究能在可重現、可對比的條件下持續進步。

6 實務檢核清單

本研究的經驗顯示，在低資源語言 ASR 的開發過程中，系統性能的崩潰往往源自流程設計與資料管理中的細節錯誤，而非模型能力的不足。

為降低此類失效發生的機率，我們整理出四個層面的實務檢核準則，期能作為後續研究與競賽實作的參考依據。

6.1 語料與對齊層面

(1) 資料一致性檢查

在導入任何語料前，應檢查其版本、來源與編碼格式是否與訓練目標一致。特別是在競賽或多階段任務中，必須確認訓練集與測試集互不重疊，以避免誤用前一階段的資料，造成結果偏差。

(2) 檔名與索引驗證

建議於預處理階段建立雜湊（hash）或索引比對機制，以確保音檔與文字標註之間的對應正確。若提交清單由自動化程式產生，應於輸出前進行隨機抽樣檢查，以防止對齊錯誤。

(3) 多語與腔調標註

在處理多腔調資料時，應明確標示腔別資訊（如大埔、詔安），並於訓練過程中進行條件化控制，以降低模型混淆不同聲學分布的風險。

6.2 模型設定與推論層面

(1) 語言提示一致性

多語模型在推論階段應確認 language token 與實際語音語言相符。若使用 Whisper 或類似架構，建議關閉固定語言參數（forced_decoder_ids = None），或透過自動語言識別（Language ID）機制動態調整，以避免模型受到錯誤語言提示的影響。

(2) 長語音處理策略

推論階段不應僅取音檔首段進行辨識。建議採用滑動視窗或重疊解碼策略，並於後處理階段進行片段合併與時間序校正，以確保長語音內容的完整性與準確性。

(3) 中途監控與早期警示

在訓練與推論過程中應設置開發集（dev set）或樣本監控機制，以即時偵測模型異常輸出。若模型於早期階段出現錯誤語言（如英語、日語或亂碼符號），應立即中止訓練並檢查語言提示與標註格式是否一致。

6.3 評估與報告層面

首先，應以 CER 作為主要評估指標，並輔以 SER 或 WER。對於非分詞語言（如中文、客語），建議採用字元層級的 CER 作為主要報告基準；若同時呈現 SER 或 WER，則須明確說明所使用的分詞策略與音節對應準則，以確保結果具可比性。

其次，建議提升錯誤分析的透明度。除了總體指標外，應附上具代表性的錯誤樣本，以呈現模型在替換、刪除與插入三類錯誤中的具體表現。此舉有助於揭示模型偏誤來源，並為後續改進提供依據。

最後，應強化評估指標的可重現性。研究者需在報告中標明所使用的評估腳本、版本與計算公式，使不同團隊得以重現並對照結果，避免因工具差異而造成評估偏差。

6.4 流程治理與版本控管

(1) 版本追蹤與紀錄保存：

所有訓練、推論與評估流程皆應透過 Git 或同等工具進行版本控管，並完整記錄關鍵參數，如 random seed、套件版本及執行命令列設定，以確保實驗可重現性。

(2) 自動化日誌與錯誤追蹤：

應建立自動化日誌系統以追蹤模型訓練過程與評估結果，並在出現異常時能快速回溯與定位，降低流程錯誤造成的資訊遺失風險。

(3) 流程文件化與可移植性：

建議將整體實驗流程以 Notebook 或 Shell 腳本形式保存，確保研究結果可由他人重現、驗證或延伸，促進低資源語言社群的資料共享與協作。

本研究總結的檢核原則顯示，低資源語言 ASR 的成功關鍵不僅在於模型選擇，更取決於語料治理與實驗流程的可驗證性。

7 結論

本研究以參與 ROCLING 2025 客語語音辨識競賽為案例，探討多語預訓練模型（Whisper）於低資源語言情境下的實務挑戰與失效機制。雖然最終結果顯示系統在決賽中完全失效（CER/SER 皆達 100%），但此極端結果反而揭示了影響低資源

語音辨識的關鍵因素：資料一致性、語言提示設定、長語音處理與流程治理。

研究過程中，我們發現語料誤用、語言提示錯置及缺乏對齊驗證等問題，足以使模型在語音仍可辨識的情況下輸出全錯結果。這說明在低資源語言任務中，流程錯誤的影響程度可遠超過模型本身的效能差距。因此，建立嚴謹的資料與流程檢核機制，比單純調整超參數或擴充模型規模更能有效提升系統穩定性與可靠性。

根據錯誤分析與實務反思，我們提出四項改進方向：
(1) 制定語料與訓練資料版本管理流程；
(2) 建立語言提示與標註一致性檢核；
(3) 引入長語音重疊解碼與分段策略；
(4) 強化評估可重現性與錯誤追蹤機制。

這些原則可作為後續客語 ASR 系統與其他低資源語言研究的基礎。

未來工作將聚焦於三個面向。首先，採用自動語言偵測結合語音特徵對齊的方法，以動態調整語言提示。其次，擴充跨腔調語料，建立能覆蓋大埔與詔安兩腔的平衡訓練集。最後，設計可視化分析工具，用以即時追蹤訓練過程中的語言漂移與對齊錯誤。

總結而言，本研究雖以「失敗案例」為出發點，但其貢獻在於提供一份可驗證、可重現、可借鑑的經驗報告，說明在低資源語音辨識領域中，「失敗」本身亦是推動技術成熟的重要養分。

References

許勝銘. (2007). 大詞彙客語語音辨識系統之初步研究 國立臺灣科技大學]. 臺灣博碩士論文知識加值系統. 台北市. https://hdl.handle.net/11296/r2d95q

羅丞邑. (2011). 以資料探勘之技術解決線上客語語音合成系統中多音字發音歧義之研究 國立中興大學]. 臺灣博碩士論文知識加值系統. 台中市. https://hdl.handle.net/11296/v5422z

吳治翰. (2012). 國語、客語及瑞典語三語言語音辨識系統之設計研究 國立中山大學]. 臺灣博碩士論文知識加值系統. 高雄市. https://hdl.handle.net/11296/aa5v2k

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023).

Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 28492–28518). Proceedings of Machine Learning Research. Available at https://proceedings.mlr.press/v202/radford23a.html

The NPTU ASR System for FSR2025 Hakka Character/Pinyin Recognition: Whisper with mBART Post-Editing and RNNLM Rescoring

Yi-Chin Huang[1]*, Yu-Heng Chen[1], Jian-Hua Wang[1], Hsiu-Chi Wu[1]
[1]Department of Computer Science and Artificial Intelligence,
National Pingtung University,
Pingtung city, Taiwan
*ychuangnptu@nptu.edu.tw

Chih-Chung Kuo[2], Chao-Shih Huang[2], Yuan-Fu Liao[2]†
[2]Speech AI Research center,
National Yang Ming Chiao Tung University,
Hsinchu city, Taiwan
†yfliao@nycu.edu.tw

摘要

本文針對 FSR-2025 客語語音辨識（Hakka Automatic Speech Recognition, ASR）競賽，統整兩個子任務：（i）客語漢字（Characters），（ii）客語拼音（Pinyin）。我們提出一條統一架構：以 Whisper [1]（大型弱標註語音辨識模型）為聲學骨幹，視情況採用 LoRA（Low-Rank Adaptation [2]）進行參數高效微調；搭配 MUSAN [3]（音樂／語音／噪音資料庫）與語速（tempo/speed）擾動 [4]之資料增強；在漢字任務中加入 mBART-50 [5,6]（多語言序列到序列模型）進行文本修正；兩個任務皆以 RNNLM [7]（Recurrent Neural Network Language Model）對 N-best 候選做重評分。在漢字任務之決賽設定下，mBART 驅動之 10-best 文本修正 + RNNLM 可達成 CER（Character Error Rate）6.26%，主辦方公布最終 CER 為 22.5%；在拼音任務中，Medium 的模型相較 Large 更適合本資料規模與腔調分布，配合 10-best RNNLM 重評分，在自訂的熱身賽測試集可得 SER（Syllable Error Rate）4.65%，最終決賽公布帶腔調成績為 14.81%。此外，我們亦分析 LID（Language Identification, 腔調辨識）在不同來源（錄製／媒體）之效益。

Abstract

This paper presents our system for the FSR-2025 Hakka Automatic Speech Recognition (ASR) Challenge, which consists of two sub-tasks: (i) Hakka Characters and (ii) Hakka Pinyin. We propose a unified architecture built upon **Whisper** [1], a large weakly supervised ASR model, as the acoustic backbone, with optional **LoRA** (Low-Rank Adaptation [2]) for parameter-efficient fine-tuning. Data augmentation techniques include the **MUSAN** [3] corpus (music/speech/noise) and **tempo/speed perturbation** [4]. For the character task, **mBART-50** [5,6], a multilingual sequence-to-sequence model, is applied for text correction, while both tasks employ an **RNNLM** [7] for N-best rescoring. Under the final evaluation setting of the character task, mBART-driven 10-best text correction combined with RNNLM rescoring achieved a **CER (Character Error Rate)** of **6.26%**, whereas the official leaderboard reported **22.5%**. For the Pinyin task, the **Medium** model proved more suitable than the **Large** model given the dataset size and accent distribution. With 10-best RNNLM rescoring, it achieved a **SER (Syllable Error Rate)** of **4.65%** on our internal warm-up test set, and the official final score (with tone information) was **14.81%**. Additionally, we analyze the contribution of **LID (Language Identification)** for accent recognition across different recording and media sources.

關鍵字：客語語音辨識、資料增強、語言模型重評分、腔調辨識、文本修正

Keywords: Hakka ASR, Data Augmentation, RNNLM N-best Rescoring, Accent Identification, Text Correction

1 緒論

客語具多腔調特性，語音資源相對稀少且口語變異度高，使語音辨識等任務的建模面臨資料稀疏與跨腔穩健性的挑戰。近年大型自監督/弱標註模型（如 Whisper [1]）展現強健跨語言能力；同時，機器翻譯領域的多語序

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 518–522
November 20-22, 2025 ©2025 Association for Computational Linguistics

列到序列模型（如 mBART-50 [5,6]）在文本層面具備校正與樣式歸一化的潛力。基於此架構，本次競賽中，我們主要設計的語音辨識系統，採取「聲學候選 → 文本修正 → 語言模型重評分」的流程，以兼顧資料規模限制與跨腔調需求。

而客語拼音 ASR 面臨腔調差異（大埔、詔安）與資料來源多樣（錄製/媒體）的雙重挑戰。大型弱標註模型 Whisper 具跨語言泛化力，但在高噪環境與腔調失配時仍需資料增強與語言模型輔助；同時，針對「不揭露腔調」的決賽情境，如何在不犧牲準確度的前提下進行腔調亦是本次競賽重點。

本系統之主要特色為：a) 針對客語漢字任務中，建立一套客語端對端 ASR + 文本修正 + LM 重評分之語音辨識系統；系統性比較熱身賽與決賽兩階段的資料切分/訓練差異；加入 mBART 後對於模型的影響。b)在客語拼音任務中，建立以 Whisper-Medium 為本的拼音 ASR 管線，實證其在本任務中優於 large-v3+LoRA；定量分析 MUSAN 噪音增強與 RNNLM 10-best 重評分的疊加效益；提出 avg_logprob 與 辭典查詢兩種 LID，並比較其在異質語料的效果；

2 資料與設定

2.1 原始語料

原始資料涵蓋大埔/詔安腔、男女聲，總計 123 位說話者、27,349 句、約 62.0 小時。本文後續實驗所用的統計與切分均依競賽規範與團隊內部規劃進行。決賽公開集 4,563 句 / 約 10.0 小時。

2.2 資料前處理

針對語料中的資料格式，為了單純處理可用之語料，若 CSV 欄位「備註」含「正確讀音」則刪除該筆；合音字的部分，將其移除星號（例：「來*去」→「來去」）。

熱身賽的語料處理，每個子語料挑 6 位說話者，隨機挑選 3 位語者之資料作為測試集、3 位作驗證集。整體統計：訓練集中包含 99 位/47.28 小時、驗證集中包含 12 位/7.10 小時、測試集中，包含 12 位/6.02 小時。而針對決賽的語料處理，為了最大化訓練的語料數，僅保留熱身語料的 5% 為測試，其餘與訓練集合併；另含媒體語料。總計：69.35 小時做為訓練集、0.50 小時的資料為驗證集；統計合計 69.39 小時。

資料增強的步驟中，我們採用 MUSAN 資料集[3]之語音（noise/speech/music）進行訓練模型的語料擴增，每段語音隨機混入 2–3 段樣本，SNR ∈ {5,10,15} dB；並以 0.5 倍速為間隔進行語速調整：慢速為原始語料的 0.7 至 0.95 倍，快速則為 1.05 至 1.5 倍。最終資料量為原始的 6 倍，但內容保持不變。在漢字任務中，將訓練語料在資料增強後，擴增前後的驗證集，在 large-v3+LoRA 的設定下，其 CER 由 8.31% 降至 1.98%，有顯著的下降。

而在拼音任務中，我們發現到以實際訓練語料的測試集的結果比較，顯示 Whisper-Medium 優於 Large +LoRA，因此後續一律以 Medium 為基底；並採用典型的訓練超參數：epoch = 5、batch = 4、lr = 1e-5、grad-accum = 2、warmup = 100。而加入語料擴增之後，實驗結果顯示，以原始訓練語料訓練的模型在乾淨語料上表現良好，但在含有雜音的語料上辨識效果明顯下降。相較之下，經過資料增強的模型在面對含雜音的語料時，確實能提升辨識表現(大埔 7.75% vs. 5.63%，詔安 9.28% vs 7.80%)。故後續訓練模型均以經過資料增強的語料進行。

3 模型定義與系統流程

3.1 Whisper + LoRA（聲學模型）

在客語漢字的任務中，我們透過實驗決定以 Whisper Large 為基底，其採 128 Mel bins 與擴增語言標記。訓練的步驟，採 LoRA 進行參數高效微調。主要超參數差異如下：熱身賽：epoch = 4、batch size = 4、grad_accum = 2、grad_checkpointing = False；決賽：epoch = 10、batch size = 16、grad_accum = 1、grad_checkpointing = True；共同設定則包含使用 AdamW、fp16、lr = 1e-4、warmup_steps = 1000；LoRA： r = 8、alpha=16、dropout=0.1、目標模組 {k/q/v/out_proj, fc1, fc2}。在拼音任務中，在實際訓練語料在測試集的結果比較，顯示 Whisper-Medium 在本任務中，優於 Large +LoRA（大埔 6.5% vs 6.98%，詔安 5.99% vs 10.12%）

3.2 RNNLM（語言模型）

為了提升辨識的效能，我們在客語漢字和拼音的任務中都加入了語言模型的後處理，希望能夠提升辨識的正確率。因此，我們總共收集了多個來源的文字作為語言模型的訓練，其中包含共 109,487 句，而來源包含教育部客語辭典例句、哈客網路學院[13]教材/試題與原始訓練文本。模型為 2-layer LSTM [8]（emb=512、hid=1024、dropout=0.3），訓練超參數如下：seq_len=256、batch=128、epochs 200、lr 2e-3（AdamW, β=0.9/0.98）、clip_grad 1.0。推論的部分，以 beam=5、LM 權重 0.5 作重評分。

而在拼音的任務中，我們更進一步地進行分析，以 LSTM/GRU [9], 2 層，embed=512，hidden=1024 訓練字/子詞級 LM，對 ASR 輸出結果 10 個候選句進行重評分。大埔腔在 beam=10、LM weight=0.5 下，WER 由 5.65% 分別下降至 3.71%（LSTM）/ 3.37%（GRU）。因此，在拼音任務中我們採用 GRU 作為語言模型。

3.3 mBART 文本修正

在決賽中，為了進一步改善漢字任務的辨識效果，我們加入了預訓練的大型語言模型的微調，希望能夠進一步地改善辨識後的成效，在此我們採用了支援多語的 mBART[5]作為基礎模型，再透過客語的文本語料來微調後，使其有能力進行更正輸入字串的下游任務。

訓練資料來自我們所訓練出來的三種聲學模型的輸出（大埔、詔安、混合）模型推理之 5 個最佳的候選句，以採樣設定（top-k=8、top-p=0.96、temperature=0.7）產生多樣候選，構成（input, target）對。接著透過同樣客語文字資料集來對 BERT-base-Chinese 進行斷詞。斷詞器是基於 bert-base-chinese 模型，經過 10 次 epoch 訓練，其超參數 batch size = 16、lr = 2e-5。訓練的語料則為主辦方所提供之兩個腔調訓練語料。斷詞的資訊則是透過台灣客語語料庫所建置的斷詞系統，其標註的詞性標記總共 18 類[12]。

我們比較斷詞前後的語料來進行文本修正，其訓練集包含了大埔腔、詔安腔與混合腔調三種模型分別進行推理，並透過 sampling 機制（top-k=8、top-p=0.96、temperature-0.7）為每段語音生成 5 個候選句，而訓練目標則為真實的原始句子。訓練總數約有 37,407 個音檔，測試集則隨機挑選 1,000 句，結果顯示，測試集中未斷詞的更正文本後，CER 為 24.77%；加入斷詞資訊後，則可改善至 19.29%，因此我們最後採用斷詞後的語料來進行訓練/驗證。mBART 訓練採用 facebook/mbart-large-50，其超參數設定如下：beam=5、lr=5e-5、batch=16、epoch=5、weight decay=0.01。另在 CE loss 上加入 Over-Edit Penalty 抑制過度修改，透過比較不同的加權值 α 後，透過驗證集決定其權重，最終尋得最佳 0.8（驗證語料中，獲得 CER 12.91%、句錯率 SER 20.83%）。

4 系統效能分析與競賽結果討論

在本節中，我們將針對熱身賽以及決賽所使用的模型以及如何推論出最終繳交的答案進行說明。

4.1 客語漢字

針對漢字任務，在熱身賽時，我們所採取的步驟主要為資料增強 → Whisper Large-LoRA → RNNLM 重評分，這三個步驟，所獲得的結果，我們自行測試的結果為混合腔調 Large 的 9.34% CER，Medium 的 10.39%。最終，主辦方公布為 8.84%，顯示在大型的育訓練模型進行模型微調後，在較複雜的客語漢字任務中較合適。此外，再加上語言模型針對 Whisper 輸出進行重新評分後，進一步提升其效能，比 baseline 的 10.42%進一步提升。

而在決賽中，我們進一步加入 mBART 的結果進行分析，因此在推論的流程中，修改為以下的方式：資料增強 → Whisper 產生 5 個候選 → mBART 文本修正並擴增 5 個候選（共 10 個候選句）→ RNNLM 打分挑選最佳解。在此，我們自行內部用熱身賽語料的 5%進行測試時，能夠從單純僅用訓練集的 CER 為 9.20% 下降至 6.26%。最終主辦方公布的結果，則是意外的有落差(22.50%)。

4.1.1 錯誤分析與改進方向

由於在決賽公布結果後，漢字的任務效果低於預期，在詳細檢測結果後發現 mBART 模型中，有預存的 token 辭典，若是沒有看過的 token，皆使用<unk>取代，這將大幅影響我們

文本修正的效能，我們統計了熱身賽與決賽中有多少 token 會受此影響，發現分別為 3,163 與 4,134 個 token 都被強制更換成<unk>。而文本修正的模型會直接忽略這些<unk>而導致最終影響辨識客語漢字的結果。若這些 token 都能被正確辨識的狀況下，我們賽後測試 Byte-fallback 後，會提升約 5%的正確率(從自行測試的 22.97%降到 17.54%)，但其效果跟主辦方公布的 baseline (17.13%)還是十分相近，可見還是有很大的進步空間。

錯誤主要的面向，一方面可能是斷詞的效果在辨識錯誤的前提下，無法將其修正，第二就是訓練文本修正的模型還不夠完善，可採用 無需固定詞彙表的序列生成模型，例如 ByT5 [10]或 CANINE [11]，其以 byte-level 或 character-level 方式建模，能夠處理 <unk> token 之問題。另外，也有研究是採取 專用的文本修正與重評分策略。以 ByT5 或 T5 為基礎的序列到序列後編輯模型，可針對 ASR 輸出中出現 <unk> 的位置進行上下文預測修補；同時，於重評分階段導入 Transformer-LM [13] 以取代傳統 RNNLM，並在評分函式中加入 <unk> 懲罰項，可進一步減少未知詞候選被選中的機率。

4.2 客語拼音

在客語拼音的任務中，在熱身賽的階段，我們分別對不同的腔調進行各自模型與混和模型的訓練與測試。結果顯示，大埔腔與詔安腔，在原始的語料中，經過資料增強加入噪音後，的確有增加其強健性，大埔腔對於熱身賽的測試從 7.75% 降至 5.63%；詔安腔亦從 9.28% 降至 7.80%（在 Whisper Medium 的設定下）。進一步加入拼音的 RNNLM 熱身賽的大埔腔測試語料從 5.65%降至 3.37%。因此，在熱身賽的階段，我們最終繳交的版本便是採用各自腔調的聲學模型，再加上語言模型的後處理的結果。最終，包含兩個腔調語料的辨識結果，主辦方公布為 13.44%。

4.2.1 個別腔調進行判斷

由於本次競賽中具有兩種腔調，分別為大埔腔跟詔安腔，其拼音的音節組成與音調的調號有些許差異，因此可能導致辨識效果不穩定，再加上決賽的語料並不會提供腔調的標籤，因此，在此我們測試了混和訓練模型跟個別訓練模型，觀察各自的現象。其中，混合腔調的模型基本上就是採兩個腔調的語料子集合進行合併後訓練，並同時比較 Whisper Large 跟 Medium 的差異。結果顯示混合模型在 large 的狀況下，其拼音的 WER 分別為 13.06% (大埔)以及 19.76% (詔安)；而 medium 的狀況下則是 5.64% (大埔) 以及 8.80% (詔安)。雖然合併語料訓練 medium 較 Large 可獲得較佳的辨識率，但比腔調單獨訓練的模型來說差異並不大，因此最終仍以腔調分開之 medium 模型作為最終模型。

為了能達到最佳的個別腔調的判斷，以符合決賽的需求，因此我們提出了兩組分辨腔調的方法。第一個方法相對單純，採用 Whisper 模型在進行辨識任務時，所輸出拼音序列的平均機率 (avg_logprob) 的數值作為模型選定。並比較兩個腔調的數值大小來做為選擇腔調的依據，若兩者數據相同時，根據前述的實驗結果看起來，混合模型有一定的準確度，因此便採用混合模型來處理此狀況。我們使用熱身賽的所有語料進行腔調的辨識，結果得出，詔安腔的正確率為 93.95%、大埔腔的正確率為 97.89%，且透過腔調辨識後，使用其辨識模型所辨識出來的拼音音節錯誤率分別為 11.67%(大埔)和 19.53%。其中雖然大埔腔的效果相較直接使用混合模型的表現好(15.89%)，但詔安腔的混合模型反而較進行腔調辨識後的結果來得較好(17.27%)，因此，此方式無法保證在未知腔調的狀態下，挑選各自腔調或是混合模型來得好。

為了得到更穩定的腔調辨識結果，以幫助挑選到合適的模型，我們希望透過兩個腔調的漢字常用辭典的方式來幫助挑選。因此，我們將在漢字任務中辨識出來的結果，透過一個預先訓練好的客文字斷詞器將其斷詞，接著透過查詢教育部台灣客語辭典[12]中，其提供的大埔腔以及詔安腔的辭典，分析斷詞後的字詞對應到兩個腔調的辭典中，何者的

測試語料	**Large Hybrid**	**Medium hybrid**	**Medium Accent**	**Best**
大埔腔 (5 hr)	28.25%	15.89%	**13.50%**	**10.70%**
詔安腔 (5 hr)	31.68%	17.27%	**16.48%**	**16.39%**

表格 1: 熱身賽拼音任務的個別腔調與混合模型結果分析.

比例較高，進而採用該腔調辨識模型來進行辨識。表格 1 則為我們透過熱身賽的語料進行測試的結果，總共比較三種模型，分別是 Whisper Large 的混合腔調模型、Whisper Medium 混合腔調模型、以及透過辭典輔助的腔調辨識後，在以該腔調辨識 Whisper Medium 的結果；此外，我們亦將使用腔調標記所辨識出來的結果(以 Best 標註)放入表中以做為參考。從結果可得知，透過斷詞再加辭典的方式，可以在詔安腔獲得近似最佳的結果，而大埔腔雖然沒有像詔安腔如此明顯，但也是目前最佳的效果，因此我們便採用此方案來辨識決賽的客語拼音任務。為了進一步增加訓練語料以改善辨識效率，在拼音的決賽模型中，我們保留少量的大埔與詔安腔的語料(各 400 筆錄製語料和 100 筆媒體語料)做為測試集，其餘皆放入訓練之中。最終拿此測試集去測試我們的拼音模型任務的結果為 5.87% (medium)，以及 5.43% (medium +腔調辨識)，並作為最終繳交的模型，主辦方公布決賽的 WER 為 14.81%。

5 結論

本文針對 FSR-2025 客語 ASR 兩子任務提出一條統一且可擴充的辨識流程：以 Whisper 作為聲學骨幹，配合 MUSAN 與語速擾動提升雜訊與腔調下的強健性；於漢字任務中加入 mBART-50 文本後編輯；兩任務皆以 RNNLM 進行 10-best 重評分。實驗顯示：在拼音任務上，Whisper Medium 較 Large-v3+LoRA 更適合本競賽資料規模與腔調分布，透過 GRU-LM 重評分可顯著降低 WER；在漢字任務上，mBART-50 可進一步降低內部評測的 CER，但最終決賽表現受限於字彙覆蓋與 <unk> 之影響。錯誤分析指出，未知詞的處理為主要瓶頸。我們的改進方向包括：採用 byte/character-level 後編輯模型（如 ByT5、CANINE）以消弭 <unk>，以 Transformer-LM 取代傳統 RNNLM 進行重評分並加入 <unk> 懲罰。整體而言，本系統所提出的模組化流程在低資源、多腔調場景具備良好可遷移性，亦為未來客語 ASR 的系統化改良提供可重現的基準。

致謝

本研究承蒙國立陽明交通大學人工智慧語音研發中心與產學創新研究學院之支持與協助，在此謹致謝忱。

參考文獻

[1] Alec Radford, Jong Wook Kim, Tao Xu, et al. 2023. Robust Speech Recognition via Large-Scale Weak Supervision (Whisper). *arXiv:2212.04356*.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:2106.09685*.

[3] Daniel Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. *Proc. Interspeech*.

[4] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio Augmentation for Speech Recognition via Speed Perturbation. *Proc. Interspeech*.

[5] Yinhan Liu, Jiatao Gu, Naman Goyal, et al. 2020. Multilingual Denoising Pre-training for Sequence-to-Sequence. *ACL* (mBART).

[6] Yuqing Tang, Chau Tran, Xian Li, et al. 2020. Multilingual Translation with Extensible Multilingual Pre-training and Finetuning. *arXiv:2008.00401* (mBART-50).

[7] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. *Interspeech*.

[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*.

[9] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, et al. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation (GRU). *EMNLP*.

[10] Linting Xue, Noah Constant, Adam Roberts, et al. 2022. ByT5: Towards a Token-Free Future with Byte-Level Models. *TACL*.

[11] Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *TACL*.

[12] 教育部臺灣客家語常用詞辭典（線上資源）。網站：Ministry of Education, Taiwan Hakka Dictionary (accessed 2025).

[13] 哈客網路學院（線上教材/試題資源，作為文本來源與詞表參考）。Website: Hakka Online Academy (accessed 2025).

Author Index