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

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：

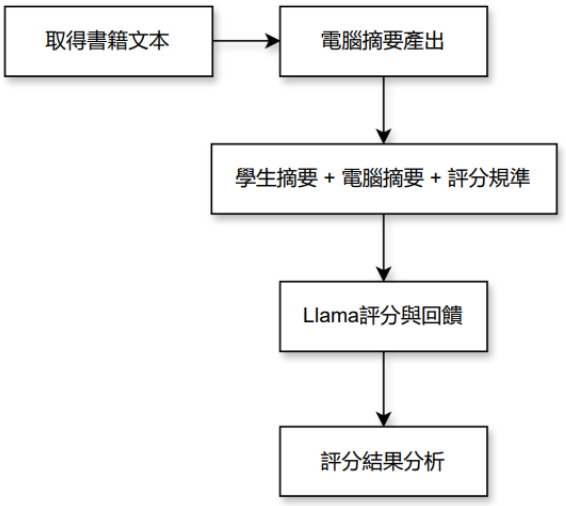

圖 1：基於 Meta-Llama-3.1-70B 產生電腦摘要，並根據此電腦摘要與評分規準對學生摘要進行評分、回饋及結果分析之實驗流程圖

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。

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「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。

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惠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「平窮有錢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，全世界至少之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。戰爭會讓在人無法繼續中人故意埋設炸彈在田裡讓在人無法繼續種田的中。	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舖直述，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