Exploring the Applicability of Large Language Model and Rubric-Based Automatic Assessment of text summaries by high school students
大型語言模型結合評分規準於高中學生文本摘要自動批改之適用性研究

黃如意 Ru-Yi Huang
國立臺灣科技大學數位學習與教育研究所
ruyi31003@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

摘要寫作訓練學生的高層次思考，反映學生從理解到闡述的思考過程。本研究旨在探討大型語言模型結合電腦摘要與評分規準的自動批改系統，在教學現場上，對高中學生於不同類型文本摘要的回饋與評分，是否與真人教師的評分與回饋方向相符。研究結果顯示，系統批改結果與真人教師的批改結果於文學性文本的總分 pearson 相關是 0.83、科普性文本的總分 pearson 相關是 0.89，皆呈現高度相關(>.7)。且系統能詳細分析學生摘要內容，給予立即性的回饋與修正指引。

Abstract

Summary writing trains students' high-level thinking and reflects their cognitive process from comprehension to elaboration. The purpose of this study is to explore whether an automated grading system—which integrates large language models with computer-based summarization and scoring rubrics—provides feedback and grading for high school students' summaries of different text types that align with the evaluation and feedback directions of human teachers.

The research findings indicate a high correlation (> .7) between the automated system's grading results and those of human teachers, with a Pearson correlation coefficient of 0.83 for the total score of literary texts and 0.89 for popular science texts. Furthermore, the system is capable of detailed analysis of students' summary content, providing immediate feedback and guidance for revision.

關鍵字：摘要自動批改、大型語言模型、評分規準、摘要教學
Keywords:Automatic abstract grading, large language models, grading standards, abstract instruction.

1 緒論

資訊科技發展快速，數位時代下成長的青少年，生活中的娛樂與社交方式大受影響。短推文與短影音平台興起，不停推送文章與影片的設計，容易讓使用者沉迷。短影音成癮 (Short video addiction, SVA) 即為過度使用短影音平台的現象，已是青少年族群當中一種普遍存在的問題(Yao et al., 2025)。

短影音成癮會導致注意力不集中、專注時間縮短、難以集中精神處理干擾訊息等認知問題(Chen et al., 2023；Nguyen et al., 2025)。這樣的認知障礙會影響學習的成效，習慣短推文與短影音的學生，接觸內容較長的文本、較艱澀且需要花時間理解的學科知識時，便容易產生認知負荷。認知負荷是干擾問題解決過程中學習的重要因素(Sweller, 1988)。Cezar 與 Maçada(2023)指出，大數據環境中，隨著認知處理負荷的增加，更有可能遇到高水平的焦慮並從事規避行為，間接地促成了認知疲勞。在教學現場的觀察印證了認知疲勞的現象，學生課堂上專注時間短暫、對需要深入思考的課題感到不耐煩，寫作和表達內容流於表面，敘述經常不夠精準。

面對現今時代的樣貌與挑戰，課綱不斷進行改正與調整，自 111 學年度起，大學入學學科能力測驗（學測）的國文考科拆分為「國語文綜合能力測驗」（國綜）與「國語文寫作能力測驗」（國寫）兩卷，評量學生閱讀

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 48–58
November 6-7, 2026 ©2026 Association for Computational Linguistics

理解及寫作能力，可見培養閱讀理解與寫作能力是當今國語文教育的重點。

因此，文本摘要與寫作的訓練，實是語文教育的基礎能力之一，摘要寫作的訓練過程中，需要學生從閱讀文章、融會貫通、到闡述表達，因應現今教育培養閱讀理解素養的教學目標。摘要(Summary)是讀者在閱讀過程中經常使用的理解策略(Duke & Pearson, 2009)。學生理解文本材料之後，吸收理解並且整合成自己的觀點。摘要學習需長時間建構與培養（馮永敏等人，2015）。但教學現場所提供的閱讀理解與摘要策略之指導顯然不足（林鳳雯、崔夢萍，2011）。課文大意教學因礙於有限的授課節數，教師難以給予學生摘要個別化的回饋（李淑華、洪碧霞，2011）。為培養數位時代的學生摘要寫作能力，以促進閱讀與寫作素養，需要克服教學時間有限、學生能力與動機薄弱、無法立即回饋等狀況。因此，教學融入工具輔助已刻不容緩。

因應上述困境，摘要寫作的研究已將電腦科技嵌入教育中(Abdi et al., 2016; Idris et al., 2011; Sung et al., 2016; Wade-Stein & Kintsch, 2004)。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(Huang et al., 2025)。這類系統強調自動化並分析學生摘要內容，並給予即時性回饋。過去，潛在語意分析法(Latent Semantic Analysis, LSA)常被用在自動批改系統當中(Kintsch et al., 2000；Landauer et al., 1998；Sung et al., 2016)，LSA為基於奇異值分解(Singular value decomposition, SVD)的技術，在產生語義空間後，其中單字、句子和整個文字可以投射至語義空間後本以向量作為表示。這些向量彼此之間的相關程度是透過它們之間的餘弦來衡量語意是否相近。使用這個餘弦測量來計算向作者提供哪些回饋(Kintsch et al., 2000)。Landauer 等人(1998)提出多項關於 LSA 的實驗，LSA 能從表面詞彙關係中抽取潛在語意，不僅能有效模擬人類的詞彙理解、知識獲得與閱讀理解歷程，也在資訊檢索、作文評分、閱讀評量及教育科技等領域展現良好成效(Landauer et al., 1998)。Kintsch 等人(2000)於在 1997 至 1998 年，使用「State the Essence」，該系統利用潛在語意分析(LSA)技術來評估學生的摘要內容，並提供即時回饋，Kintsch 等人在 1999 年對軟體進行了重大改版，修正回饋方式與介面，重新定名為「Summary Street」(Kintsch et al., 2000)。該系統可回饋學生是否涵蓋核心概念、內容長度是否適切，也告訴學生摘要中缺少哪些訊息，提供對冗餘、無關內容的評論回饋 (Kintsch et al., 2000)。在中文環境中，馮樹仁(2001)發展國小摘要寫作評量系統，該實驗探討以潛在性語意分析的方法所設計的評量系統是否具有評量學生摘要寫作的效果，其成就測驗與各項寫作評量成績之間的效標關聯效度，大部分皆達到顯著水準。張國恩與宋曜廷(2005)利用 LSA 技術建立兩個主題（族群與群落、端午節）的語意空間，結果發現以向量餘弦值為標準的評分方式，與教師評閱的分數之間有不錯的相關。

潛在語意分析(LSA)用於摘要自動批改系統中，固然有其優點，但是潛在語意分析的評分仰賴語料庫及標準答案，在評分一致性上有時會稍有影響，語意理解也會影響評分結果。例如藍敏杰(2007) 探討國小生摘要兩個文本，以教師人工評量與結合 LSA 之自行開發的系統評量，評分分數是否相關，因語意空間理解不同，其中一則關於端午節主題的文本相關性達顯著，另一則關於丹頂鶴主題的文本則呈現不一致。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(Huang et al., 2025)。由此可知，潛在語意分析(LSA)在自然語言理解、語境和語序理解上的不足，會使摘要批改判定的結果不夠準確。

近年來，大型語言模型(LLM)發展快速，徹底改變了自然語言的理解和生成。他們擁有深厚的語言理解能力、類人的文本生成能力、情境意識和強大的解決問題的能力 (Yao et al., 2024)。當大型語言模型結合自動批改系統，善用其理解能力，給予專家設計的評分規準，能夠協助教師給予學生評分與回饋。過去於中文的自動批改系統研究中，曾有結合大型語言模型和評分規準，檢核國小學生書籍摘要結果，研究顯示自動批改系統可以提供具體且清晰的即時回饋，能指出摘要中遺漏的重要觀念、也能針對語構安排與語法錯誤提出修正建議 (Huang et al., 2025)。從上述的文獻可知有幾個研究的缺口：其一，自動

批改系統的研究中，以英文摘要的批改較多，中文摘要批改結果的研究相對較少；其二，中文摘要批改上，在高層次語文能力評估上仍存在限制；其三，缺少對於中文長文本的摘要批改研究；其四，缺少對於高中學生文本摘要批改結果的實證研究；最後，現存研究中，缺少對於不同類型文本摘要的研究（黃淇瀅，2020；Huang et al., 2025）。

有鑑於結合大型語言模型的自動摘要批改在實際教學現場的實驗仍不足，尤其是中文的摘要批改、各類型文本的摘要批改效果，仍存在缺口，大型語言系統是否可以如同真人教師，給予學生準確評分與回饋，尚未可知。本研究以 Google 於 2026 年 4 月發表，目前 Gemma 4 家族中體積最大、效能最強的密集模型的開源大型語言模型 Gemma 4 31B(Google AI, 2026)結合評分規準與電腦摘要，探討高中學生文本摘要自動批改，與專家老師批改結果是否具相關性？且自動批改系統給予學生的回饋是否具適切性?

本研究研究問題如下：

1. 開源大型語言模型 Gemma 4 31B 結合評分規準與電腦摘要，於高中學生文本摘要自動批改，是否能針對高中學生的中文文本摘要，給出與真人老師相符的評分與回饋?
2. 開源大型語言模型 Gemma 4 31B 結合評分規準與電腦摘要，於高中學生文本摘要自動批改，是否能針對不同類型的文本摘要，給出與真人老師相符的評分與回饋?
3. 開源大型語言模型 Gemma 4 31B 結合評分規準與電腦摘要，於高中學生文本摘要自動批改，以專家老師角度評估，針對高中學生的文本摘要給出的評分與回饋，對學生而言是否有所幫助？

2 文獻探討

摘要寫作是建立對複雜材料扎實理解的一種方法，相比於常常讓學生不知所措的開放式寫作任務，摘要寫作要學生為了特定目的進行摘要，充分且簡明地表達內容，同時清晰表達自己的理解 (Kintsch et al., 2000)。摘要不是單純地增加或刪除單個詞彙、短語或句子，而是在單個句子中結合多個觀點，以及跨段落的各種細節進行概括，重組文本內容(Kintsch et al., 2000)。早期自動評分系統多以潛在語義分析(Latent Semantic Analysis, LSA)基礎，系統生成一個語意空間，在其中詞彙、句子和整個文本都可以表示為向量(Kintsch et al., 2000)。這些向量彼此之間的關聯緊密程度是透過它們之間的餘弦值相似度(cosine similarity)來測量語意的相似度，以比對語意空間中的關鍵字與摘要內容的關鍵字，以回應的準確與否，決定後續回饋與修正的內容 (Kintsch et al., 2000)。

Kintsch 等人在 1997-1998 年間，以兩個由協同教學組成的班級參與了使用早期版本摘要工具「State the Essence」的實驗，該系統旨在支援學生在三個課程單元中的摘要寫作，學生首先提前使用文書處理器或紙筆撰寫好摘要，然後將摘要貼上或輸入到 State the Essence 中，以獲得如何修改的反饋。在此實驗中，學生樂於挑戰以提高他們的分數，並能保持對任務的專注。然而，分數並不總是可靠的，同時，學生被大量的反饋所壓倒，對要處理的多重問題感到沮喪。許多學生需要有非常明確的指導，並學習有意義的修改摘要內容 (Kintsch et al., 2000)。

Kintsch 等人在 1999 年修改系統介面、分數回饋形式、錯字與待修改處提示方式，重新命名為「Summary Street」，兩個六年級班級的學生在反饋的引導下，使用 Summary Street 在線上撰寫或修改他們的摘要，該摘要工具透過幫助學生自己識別其寫作的特定問題，給予學生更多的寫作實踐與反饋(Kintsch et al., 2000)。

在中文環境中，馮樹仁(2001)發展國小摘要寫作評量系統。評量的系統主要參考「潛在語意分析」的摘要寫作評量的方法，從教材及學生的摘要作品中，找出關鍵詞做為評量的依據，以台中縣富春國小六年級的 79 名學生為實驗的對象。在經過摘要寫作及心得寫作的實驗過程及單元成就測驗後，評估透過評量系統所獲得的評量指標與兩位教師評量結果和學生成就測驗之間的相關性。此研究中發展的系統能以關鍵詞為摘要寫作評量的依據，但關鍵詞的擷取部分仍然需要人工的作業配合，且此系統也尚未能直接告訴同學摘要作品中多餘的詞句、內容重覆的地方或是目前所疏忽的摘要重點等訊息(馮樹仁，2001)。

藍敏杰(2007) 發展摘要寫作系統，探討「特定領域文章 (P)」、「關鍵詞 (K)」及「奇

異值保留個數 (S)」三項因素對建構語意空間及評量效度之影響，以二重國小五年級的 107 名學生為實驗對象。在經過「端午節」主題和「丹頂鶴」主題兩篇摘要寫作測驗及相關性分析後，得到以下結論：兩個實驗之分析結果呈現 LSA 的評分在端午節文章中與兩位教師人工評量相關性皆達顯著水準；但在丹頂鶴文章中，因為語意空間對文本的理解與教師不同，評分則呈現不一致現象。因此，以 LSA 作為摘要評量分數的效度，在真人教師與系統之間，以及在不同文體的理解尚未有一致結論。

LSA 受限於語意空間，也依賴對「詞彙與上下文」矩陣的奇異值分解(SVD)(Landauer et al., 1997)。在語序不同狀況下的語義理解，有時與人類仍存在差異。Landauer 等人(1997)曾舉例，以下兩個句子對 LSA 來說是完全相同的，但對人類讀者來說卻有非常不同的含意："It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem."（那天酗酒的不是銷售經理，而是那個有嚴重飲酒問題的辦公室職員。）"That day the office manager, who was drinking, hit the problem sales worker with a bottle, but it was not serious."（那天，正在喝酒的辦公室經理用瓶子打了那個有問題的銷售人員，但並不嚴重。）。過去潛在語意分析依賴語料庫與標準答案，分析學生答案與標準答案相似程度，為摘要內容評分，忽略語詞相近但語序不同，所產生的語義差異，在評分與回饋上產生評量漏洞。因此，在摘要自動批改系統的發展中，希望追求更貼近人類的語義理解。當機器學習在不同語境當中，詞彙排列與運用的準確程度能夠更近似於人類語言，批改上的準確率就能更優良。

近年來，受益於科技的發展與人工智慧急遽的進步，大型語言模型(LLM)的演進推動了人工智能(AI)的變革性轉變(Zhao et al., 2026)。透過規模自我監督訓練、架構創新和數據策劃策略來建立核心模型能力、上下文學習(In-context learning)、提示詞工程(Prompt engineering)和代理式推理(Agentic reasoning)，以優化真實世界的部署並實現與外部環境的有效互動(Zhao et al., 2026)。自然語言處理(NLP)的最新進展促成了強大語言模型的誕生，包括 ChatGPT 等大型語言模型 (Liu et al., 2023)。

Zhang 等人(2024)以人類撰稿員和大型語言模型(LLM)的英文新聞摘要進行評量，以盲測雙人配對比較（Pairwise Comparison）考察對於人類撰稿員和 LLM 摘要的整體偏好（Overall Preference）以及資訊量偏好（Informative Preference），評審結果總計有 50.4% 傾向人類，49.6% 傾向 LLM，幾乎平手。實驗表明，LLM 的摘要表現與人類不相上下，未來給予大型語言模型的摘要更精準的指令微調，便可適用於不同情境需求中。

Liu 等人(2023) 提出 G-EVAL，這是一個利用大型語言模型結合思維鏈(CoT)和表單填寫範式(Form-filling paradigm)來評估自然語言生成(NLG)輸出品質的框架(framework)。以人類撰稿員與 GPT-3.5 摘要英文新聞。G-EVAL 與人類同時為「人類撰寫的摘要」和「GPT-3.5 產出的摘要」評分，G-EVAL 在摘要評分任務上與人類的評分 Spearman 相關係數達到了 0.514。

摘要自動批改系統結合大型語言模型，利用其對自然語言處理的優勢，能解決利用潛在語意分析時，對於文本理解有所偏誤的情況。但就算如 ChatGPT 和 GPT-4 這樣的提示型語言學習模型 (LLM) 在自然語言生成 (NLG) 評估方面表現出色，但在實踐中仍不可避免地存在一些缺陷，例如商用大型語言模型成本高昂、學習資料隱私等問題(Alzamil et al., 2025；Pan et al., 2025)。且以 G-EVAL 為例，出現人類評分認為人類摘要優於 GPT-3.5 摘要，但大型語言模型給 GPT-3.5 摘要更高分的情況 (Gao et al., 2025)。可見運用大型語言模型時，產出穩定品質的摘要、降低成本、降低使用門檻，是開發大型語言模型進行摘要評量系統，需要克服的問題。給予大型語言模型更清楚的提示詞，以保障系統可以產出穩定且內容正確的摘要、給予大型語言模型更精準的評分規準，以保障評分結果一致性。

Huang 等人(2025)提出以開源大型語言模型 Meta-Llama-3.1-70B(Meta AI, 2024)作為電腦自動摘要與評分工具，在此研究當中，針對國小生書籍摘要進行自動批改，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。現存研究中，也缺少對於不同類型長篇文本摘要在不同學習階段，如國中或高中的實證研究(黃淇瀅，2020；Huang et al., 2025）。

在中文環境中，對於開源大型語言模型結合摘要自動批改系統的開發仍為少數，更缺少在不同教學階段的實證研究。由上述文獻可知，對於開源大型語言模型結合摘要自動批改系統仍存在以下實驗缺口：其一，尚須驗證開源大型語言模型結合自動摘要批改系統的評量效果是否可比擬真人教師，驗證在中文環境中，開源大型語言模型結合評分規準是否能精確批改摘要作品；其二，針對不同類型文本，開源大型語言模型結合評分規準是否能有效評量學生摘要；其三，開源大型語言模型結合評分規準是否能給予學生相對應的修正意見。有鑑於此，本研究將基於 Huang 等人(2025)的基礎，將大型語言系統 Meta-Llama-3.1-70B 結合評分標準的摘要自動批改系統，替換成最新的開源大型語言模型 Gemma 4 31B，並選用文學性散文與科普性散文各一篇，於批改高中生摘要寫作中，檢驗與真人教師評分的 pearson 相關係數，檢視開源大型語言模型結合評分規準的自動摘要評分系統，在教學現場上的適用性。

3 研究設計

本研究基於開源大型語言模型 Gemma 4 31B結合評分規準與電腦摘要，批改高中生於文學性散文與科普性散文的摘要寫作，檢驗系統與真人教師評分的 pearson 相關係數，希望藉以驗證該系統於教學現場上，可以於摘要批改上提供近似於真人教師的評分結果，令現場教師可以使用該系統以降低教學負擔、學生可以使用該系統以取得立即性的回饋。由於高中生語文學習教材文長經常達到上千字，內容主題廣泛，學生寫作能力參差不齊、語言能力運用程度有個別化差異，運用該系統評分與回饋，以期可以達到與真人教師同等成效。

3.1 研究對象與研究工具

3.1.1 研究對象

本研究欲檢視開源大型語言模型結合評分規準的自動摘要評分系統，是否能穩定且精準給予學生評分與回饋，此一目標不受限單種類文章，因此實驗中安排學生摘要文學性文本與科普性文本，以增加系統適用性的可信度。但因受限於高中課程與時間安排，因此採兩班同步進行實驗，一個班摘要文學性文本、一個班摘要科普性文本，研究目的相同，希望探究學生摘要文本之後，系統評分與真人教師評分的相關性。

本研究以台北市某高中一年級兩個班級的學生作為研究對象，A 班級施測人數 29 人、B 班級施測人數 24 人。學生來自北部地區包含台北、新北、基隆等縣市。班級採常態編班，兩個班級從高中入學起，皆由同一位老師教導國文課程，學生程度相仿。本研究為國科會研究計畫 114-2628-H-011 -002 -MY3，所有教學實驗皆已通過倫理審查、完成知情同意書簽署及資料去識別化。

3.1.2 研究工具

本研究使用 Google 最新的開源大型語言模型 Gemma 4 31B 作為電腦自動摘要與評分工具(Google AI, 2026)。Gemma 4 31B 是由 Google 開源，為 2026 年最新進展中為最新技術的中大型輕量化開放模型，它能高度理解並嚴格執行複雜或多步驟的提示詞指令(Google AI, 2026)。

本研究透過開源大型語言模型結合評分規準所設計的「Smart writing 系統」可以讓使用者輸入摘要內容後，給予評分與回饋。本研究將比較系統評分與真人教師評分之間是否能達高度相似，以推測此系統於教學現場上的適用性。

3.2 實驗流程

本研究透過開源大型語言模型結合評分規準與電腦摘要所設計的「Smart writing系統」讓使用者輸入摘要內容後，給予評分與回饋。

首先，教師針對高中生閱讀能力與課程，選擇文學性及科普性兩類文本各一篇；接著，透過提示詞(prompt)引導模型對書籍產生摘要。因系統目標協助學生摘要書籍或長文，先引導模型緊扣所提供的評分標準摘要文章重點。一則避免每一次摘要評分接需重新讀取書籍或文章全文，拉長比對與評分時間；二則令評分有所本，避免評分過程中，系統失準使檢驗方向流於瑣碎或偏離摘要目標書籍或文

章主旨。兩班學生分別閱讀文學性文本和科普性文本，並書寫摘要；依據專家所制定的摘要評分規準，教師根據評分規準對學生摘要評分；運用Smart writing系統，根據相同的評分規準與電腦摘要對學生摘要產生自動評分與回饋。最後，計算專家老師與Smart writing系統批改之分數的pearson相關係數。本研究整體實驗流程圖如下圖1所示：

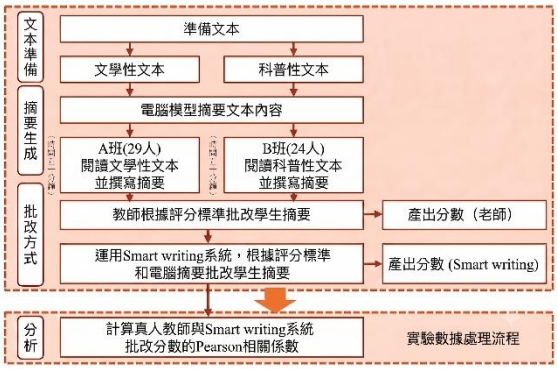

圖1、實驗流程圖，運用Smart writing系統，根據評分規準與電腦摘要批改學生摘要。

3.2.1 文本準備

本研究以中文文本供學生進行摘要，測試系統針對不同書寫方式的文本摘要批改能力。本研究分別選擇一篇文學性文本與一篇科普文本，文學性文本內容較具情意，書寫內容會有較深刻的情感描述或故事情節鋪陳、科普性文本內容較知性，書寫內容會針對某一現象或科學概念，以淺顯易懂的方式解釋。因兩類文章目標不同，一為抒發情感、一為解釋現象，兩類文章的寫作筆法會有明顯區別，本次由教師挑選兩篇可搭配學生教材課程的補充文章，其一為文學作家所寫的親情散文，共 5969 字、其二則以李白為切入對象，探討酒精對於人體影響的科普性文章，共 4344 字。

3.2.2 摘要生成：電腦模型摘要文本內容

本研究使用Gemma 4 31B作為電腦自動摘要與評分工具，temperature設定0.3、top-p設定0.2、top-k設定20。透過提示語(prompt)設計並搭配上下文感知分層合併(Context-Aware Hierarchical Mergin)引導模型對書籍產生電腦摘要(Ou & Lapata, 2025)。上下文感知分層合併(Context-Aware Hierarchical Mergin)是一種專門用於處理長文件摘要的改良技術，旨在解決大型語言模型(LLM)在處理長文本時容易產生幻覺與事實不準確的問題。在合併摘要的過程中，引入原始文件中的「相關上下文(Context)」作為輔助，為模型提供真實的事實資訊，進而減輕幻覺並提高摘要的忠實度(Ou & Lapata, 2025)。透過大型語言模型生成文章摘要，再結合評分標準，以利後續為學生撰寫之摘要進行評分與回饋。

3.2.3 摘要生成：學生撰寫摘要內容

本研究以台北市某高中一年級兩個班級的學生作為研究對象，兩個班級皆由同一位老師教導國文課程。為測驗兩類不同文本、也為增加樣本多元性，因此以兩個班為研究對象，分別閱讀文學性和科普性文本，以增加研究樣本量與多樣性。又考慮在有限的教學時間中，為不因班級課程進度規劃、學生閱讀疲乏等因素影響影響，一個班只閱讀一篇文本。A 班級施測人數 29 人、B 班級施測人數 24 人，施測時間為高一下學期學期中，請學生在 50 分鐘內閱讀指定文本，並撰寫摘要。A 班級收回 29 份摘要、B 班級收回 24 份摘要。

3.2.4 批改方法

本研究評分標準參考Huang等人(2025)在相關文獻中所整理的閱讀摘要評分標準(Chen et al., 2023；Morris et al., 2024；Özdemir, 2018)。總共分成：理解與準確性、組織結構與邏輯條理、簡潔性、語言表達與文法、以及重述能力等五大面向，並且再細分成核心概念、重要事件、理解正確、結構層次與完整性、邏輯與條理性、句段連貫性、精煉、冗贅、用字精確、修辭、語句流暢性、重述等十二項，每個項目最高五分，最低一分，總分最高六十分，用以評估學生摘要品質。真人老師也依循同樣的摘要標準進行評分。

給予系統的提示詞中，規定系統以繁體中文給予回饋，且須針對十二個面向分別給分，並且附上評分說明與總分。

4 研究結果與分析

本研究共收回 A 班級 29 份摘要結果以及 B 班級 24 份摘要結果，A 班級閱讀文本為文學性之散文文本，B班級閱讀科普性文本。學生摘要後內容，會由真人教師批改並給出分數，並且由結合開源大型語言模型 Gemma 4 31B 電腦摘要與評分規準的 Smart writing 系統批改，給出分數與回饋，再經分析比較真人教師與系統評分的 pearson 相關係數。本研究

採相關係數分析，根據 Hinkle 等人 (2003)相關係數 0.3 以下為低相關，0.3~0.7 為中等相關，0.7 以上為高度相關。

4.1 文學性文本電腦摘要與真人教師評分

如表1所示，學生摘要文學性文本後，電腦摘要評分與真人教師評分的皮爾森相關係數如下：「核心概念」的 pearson 相關是 0.58；「重要事件」的 pearson 相關是 0.84；「理解正確」的 pearson 相關是 0.67；「結構層次與完整性」的 pearson 相關是 0.59；「邏輯與條理性」的 pearson 相關是 0.63；「句段連貫性」的 pearson 相關是 0.78；「精煉」的 pearson 相關是 0.86；「冗贅」的 pearson 相關是 0.43；「用字精確」的 pearson 相關是 0.35；「修辭」的 pearson 相關是 0.37；「語句流暢性」的 pearson 相關是 0.39；「改寫」的 pearson 相關是 0.70。針對整體加總後分數，真人教師批改總分與系統批改總分的 pearson 相關是 0.83。

統計結果顯示，學生摘要文學性文本後，電腦摘要評分與真人教師評分的總分相關係數為 0.83，達到高度相關。評分細項的 12 個項目中，達高度相關項目共4項，分別為「重要事件」、「句段連貫性」、「精煉」、「改寫」，其餘8個子項目為中度相關。從次數分配來看，1/3 的評分項目達高度相關水準。

評分項目	Pearson 相關係數 (*r*)	相關程度
核心概念	0.58	中度相關
重要事件	**0.84**	**高度相關**
理解正確	0.67	中度相關
結構層次與完整性	0.59	中度相關
邏輯與條理性	0.63	中度相關
句段連貫性	**0.78**	**高度相關**
精煉	**0.86**	**高度相關**
冗贅	0.43	中度相關
用字精確	0.35	中度相關
修辭	0.37	中度相關
語句流暢性	0.39	中度相關
改寫	**0.70**	**高度相關**
整體加總總分	**0.83**	**高度相關**

表 1、文學性文本電腦摘要評分與真人教師評分的 pearson 相關係數。

4.2 科普性文本電腦摘要與真人教師評分

如表2所示，學生摘要科普性文本後，電腦摘要評分與真人教師評分的皮爾森相關係數如下：「核心概念」pearson 相關是 0.75；「重要事件」的 pearson 相關是 0.82；「理解正確」的 pearson 相關是 0.81；「結構層次與完整性」的 pearson 相關是 0.75；「邏輯與條理性」的 pearson 相關是 0.90；「句段連貫性」的 pearson 相關是 0.76；「精煉」的 pearson 相關是 0.73；「冗贅」的 pearson 相關是 0.59；「用字精確」的 pearson 相關是 0.45；「修辭」的 pearson 相關是 0.70；「語句流暢性」的 pearson 相關是 0.69；「改寫」的 pearson 相關是 0.75。針對整體加總後分數，真人教師批改總分與系統批改總分的 pearson 相關是 0.89。

統計結果顯示，學生摘要科普性文本後，電腦摘要評分與真人教師評分的總分相關係數為 0.89，達到高度相關。評分細項的 12 個項目中，達高度相關項目共9項，分別為「核心概念」、「重要事件」、「理解正確」、「結構層次與完整性」、「邏輯與條理性」、「句段連貫性」、「精煉」、「修辭」、「改寫」其餘3個子項目為中度相關。從次數分配來看，3/4 的評分項目達高度相關的水準。

評分項目	Pearson 相關係數 (*r*)	相關程度
核心概念	**0.75**	**高度相關**
重要事件	**0.82**	**高度相關**
理解正確	**0.81**	**高度相關**
結構層次與完整性	**0.75**	**高度相關**
邏輯與條理性	**0.90**	**高度相關**
句段連貫性	**0.76**	**高度相關**
精煉	**0.73**	**高度相關**
冗贅	0.59	中度相關
用字精確	0.45	中度相關
修辭	**0.70**	**高度相關**
語句流暢性	0.69	中度相關
改寫	**0.75**	**高度相關**
整體加總總分	**0.89**	**高度相關**

表 2、科普性文本電腦摘要評分與真人教師評分的 pearson 相關係數。

5 結論

開源大型語言模型 Gemma 4 31B 結合評分規準和電腦摘要的 Smart Writing 系統，在整體評分與真人老師批改結果呈現高度相關。

最後本研究針對研究問題進行綜合性的討論：

研究問題一、大型語言模型 Gemma 4 31B 結合評分規準與電腦摘要於高中學生文本摘要自動批改，是否能針對高中學生的中文文本摘要，給出與真人老師相符的評分與回饋?

研究結果顯示，系統批改結果與真人教師的批改結果，文學性文本的總分 pearson 相關是 0.83、科普性文本的總分 pearson 相關是 0.89，呈現高度相關(>.7)。大型語言模型 Gemma 4 31B 結合評分規準與電腦摘要於高中學生文本摘要自動批改能夠針對高中學生的中文文本摘要，給出與專家老師相符的評分。

研究問題二、大型語言模型 Gemma 4 31B 結合評分規準與電腦摘要於高中學生文本摘要自動批改，是否能針對不同類型的文本摘要，給出與真人老師相符的評分與回饋?

學生摘要文學性文本後，評分細項的 12 個項目中，達高度相關項目共 4 項，其餘 8 個子項目為中度相關，整體而言有 1/3 的評分項目達到高度相關的水準。電腦摘要評分與真人教師評分的總分相關係數為 0.83，達到高度相關。

學生摘要科普性文本後，評分細項的 12 個項目中，達高度相關項目共 9 項，其餘 3 個子項目為中度相關。整體而言有 3/4 的評分項目達到高度相關的水準。電腦摘要評分與真人教師評分的總分相關係數為 0.89，達到高度相關。可見大型語言模型 Gemma 4 31B 結合評分規準與電腦摘要於高中學生文本摘要自動批改，能夠針對不同類型的文本摘要，給出與真人老師相符的評分。

研究問題三、大型語言模型 Gemma 4 31B 結合評分規準與電腦摘要於高中學生文本摘要自動批改，以專家老師角度評估，針對高中學生的文本摘要給出的評分與回饋，對學生而言是否有所幫助？

以一篇於總分 60 分中取得 31 分，表現中等的文學性文本學生摘要為例，如圖 2 系統回饋所示，系統給予的回饋非常詳盡，內容並不會籠統敘述優或劣，能針對 12 個子面向分別給出具體修改方針。


總分（系統自動換算）：31/60（52分）（已抓到 12/12 個面向分數）

核心概念：§3| 回饋：你捕捉到了幾個這個核心，這是一個很好的開始。如果能將父女關係的疏離與追尋更具體地寫出來會更完整。建議可以嘗試描述作者在父親給世後的心理狀態。這樣能讓讀者更感受到文章的深層主旨。相信你只要多留意細節，一定能寫得更深刻。加油，你已經掌握了基本概念囉。

重要事件：§2| 回饋：你提到了父親去世這個關鍵轉折，這點觀察得很正確。不過文中關於童年的互動與死後的追尋過程也相當重要。如果能將這些事件依序串聯，故事會更完整。試著把文章中的時間線梳理一遍看看。這些細節能幫助你更全面地理解作者的遺憾。期待看到你更豐富的敘述。

理解正確：§3| 回饋：你理解了父親對女兒隱含的愛，這展現了你的同理心。不過關於「重男輕女」的誤方，可以描述得更精確一些。文中提到的無力感與孤獨感也是理解的重點。建議可以再深入思考作者在面對死亡時的心理衝擊。慢慢閱讀，你會發現更多細緻的情感層次。你的理解方向是對的，只要再深入一點點。

構層次與完整性：§2| 回饋：目前的摘要較像是一段隨筆，缺乏明確的起承轉合。建議可以將內容分為「生前關係」、「意外轉折」與「事後反思」三個部分。這樣能讓讀者更清晰地看出情感的推進過程。嘗試建立一個簡單的框架來組織你的想法。完整的結構能讓你的感悟更有說服力。相信你一定能組織出更清晰的篇章。

邏輯與條理性：§3| 回饋：你呈現了基本的因果關係，讓讀者知道事情發生的順序。但如果能強化「渴望」與「缺失」之間的對比，邏輯會更強。目前的敘述較為平鋪直敘，缺乏層次感。可以試著思考為什麼「沒把話說開」會導致後來的遺憾。透過邏輯的推演，能讓你的摘要更有深度。你已經有基礎了，試著挑戰更複雜的邏輯結構吧。

句段連貫性：§2| 回饋：句子之間目前的銜接較為破碎，閱讀起來略感跳躍。建議可以使用一些連續詞，例如「然而」、「因此」或「隨後」。這樣能讓讀者在閱讀時感受到情感的流動。試著將短句合併成更完整的長句，增加流暢度。運用的文字能讓你的心意傳達得更自然，多練習很棒，你的文字會變得更溫潤。

精煉：§3| 回饋：你的表達很直白，但部分資訊可以再濃縮得更精準。例如「在重男輕女的觀念中」可以與後文的愛合併討論，嘗試用一個精練的詞彙來代替較長的描述。**精煉**的文字往往能產生更強大的情感衝擊力。你可以試著刪除重複的詞彙，讓這點更突出，相信你能夠在簡潔中傳達深意。

完賞：§4| 回饋：整體而言沒有太多的離題資訊，表現得相當不錯。僅有少部分關於「說開」的重複描述可以再縮整。如果能將重複的意象合併，文章會更乾淨俐落。目前的篇幅適中，沒有過多的贅字。只要稍稍修飾重複的語句，就會非常完美。讀繼續保持這種專注於重點的寫作習慣。

用字精確：§3| 回饋：你使用了「說開」這個詞，雖然親切但略顯口語化。建議可以使用「溝通」、「化解」或「坦誠」等詞彙來提升精確度。對於情感的描述可以嘗試更多樣化的形容詞。精確的用字能讓讀者更準確地感受到作者的悲憫。試著閱讀更多文學作品來豐富你的詞庫。你的表達很真誠，加上精確的用字會更好。

修辭：§2| 回饋：目前的文字以直述為主，較少運用**修辭**技巧來強化情感。可以嘗試使用比喻或對比，來呈現父女之間那種微妙的張力。例如將遺憾比作烙印，會讓文字更具感染力。感謝的**修辭**能讓你的摘要從「記錄」變成「文學」。不要害怕嘗試新的表達方式，讓文字跳舞吧！期待看到你運用**修辭**展現出的才華。

語句流暢性：§2| 回饋：部分句子的語序較為口語，讀起來像是在對話而非摘要。建議注意標點符號的使用，避免句子過長而缺乏停頓。試著朗讀一遍你的文字，找出感覺生硬的地方。流暢的語句能讓讀者更專注於你傳達的情感。只要稍微調整語序，你的文字會變得非常自然。慢慢練習，你會發現寫作的樂趣。

改寫：§2| 回饋：目前的內容較多是直接概括，缺乏對原文語言的轉換與重構。建議嘗試用自己的話重新詮釋作者的心理狀態，而非僅是描述事件。**重述**的過程其實就是深度思考的過程。試著跳脫原文的框架，用你的視角去詮釋這段複雜。這能展現你獨立思考與語言轉換的能力。相信你一定能寫出屬於自己的深刻詮釋。

圖 2、系統回饋學生文學性文本摘要的內容。

但是，系統在對於判定高分的摘要中，則長篇的回饋裡，幾乎都在整理學生摘要內容的優點，沒辦法如同真人老師給予更多面向或加深加廣的回饋。以一篇於總分 60 分中取得 59 分的科普性文本學生摘要為例，如圖 3 系統回饋所示。「科普性」主題文章摘要，應說明文章中想要分析的科普知識，系統於各分項肯定學生的流暢敘述，卻未點出學生僅略述科普內容的缺失。此一現象可以在未來加強提示詞指引，針對不同類型文本，對於摘要寫作的總體性要求給出更詳細的規準。


總分（系統自動換算）：59/60（98分）（已抓到 12/12 個面向分數）

核心概念：§5| 回饋：你非常出色地掌握了文章的主旨，將李白的文學傾向與科學分析完美結合。文字中清晰地交代了探討物件與核心目的。展現了你敏銳的觀察力。能將複雜的內容濃縮成如此完整的句段，真的非常令人欣慰。請繼續保持這種對核心資訊的精準捕捉能力。你讓讀者能一眼看出本文的重點，這是一個很棒的開始。相信你在閱讀理解方面有著極高的天賦。

重要事件：§5| 回饋：你完整地涵蓋了文中提到的三個詩作以及酒精的生理影響，沒有遺漏任何關鍵點。從文學場景到科學機制的轉折處理得恰到好處，顯示出你閱讀時非常專注。每個重要環節都得到了妥善的呈現，讀摘要內容十分充實。看到你如此細心地梳理資訊，老師感到很驕傲。這種全面掌握資訊的能力將對你未來的學習大有幫助。請給自己一點掌聲，你做得非常好。

理解正確：§5| 回饋：你的陳述非常精確，完全正確地理解了酒精如何影響大腦以及李白的心理狀態。將「多巴胺」與「興奮感」的關係描述得十分清晰，顯示出你對科學知識的吸收力很強。文學分析的部分也精準地捕捉到了「以酒行樂」的精髓。你的理解深度讓這篇摘要具有很高的可信度。能將跨領域的知識如此正確地整合，真的非常不容易。你對文字的解讀非常到位，請繼續保持。

構層次與完整性：§5| 回饋：文章的結構層次分明，從文學現象到科學分析，最後以總結收尾，邏輯推進非常自然。每一段都有明確的分工，讓閱讀過程感到非常舒適且完整。你成功地建立了一個清晰的框架，使資訊地圖一目瞭然。這種有條理的寫作習慣能讓你的表達更具說服力。整體篇章結構完整，展現了你優秀的組織能力。這是一篇結構典範，值得你為自己感到自豪。

邏輯與條理性：§5| 回饋：資訊之間的邏輯關係非常清楚，因果與遞進的結構運用得恰到好處。你將李白的行為與生理機制之間的關聯性處理得十分順暢，條理分明。讀者能輕易跟隨你的思路，從現象深入到本質。這種清晰的邏輯思考能力是學習中最重要的資產之一。你將複雜的交叉視角與化為易懂的邏輯鏈條，表現得非常優秀。請繼續運用這種理性的思考方式來面對挑戰。

句段連貫性：§5| 回饋：句與句之間的銜接非常自然，段落轉折流暢，閱讀起來毫無阻礙。你巧妙地運用了連線詞，使整篇文章像流水般順暢。整體連貫性極佳，展現了你對語言掌控的細膩程度。這種流暢的閱讀體驗是來自於你對文字的用心經營。看到你如此注重細節的自然度，老師感到很欣慰。你的文字具有一種溫潤的節奏感，請繼續保持。

精煉：§5| 回饋：你能以極其簡潔的文字傳達核心資訊，有效濃縮了原文的精華。沒有贅字地將複雜的科學過程簡化，且不失準確性，這非常困難。每句話都簡潔有力，直接切中要害，展現了極高的濃縮能力。你成功地在精簡與完整之間找到了完美的平衡點。這種**精煉**的表達方式能讓溝通效率大大提升。你的文字功底非常紮實，表現得十分出色。

完賞：§5| 回饋：全文幾乎沒有**完賞**的語句，資訊聚焦度極高，完全沒有離題的情況。你精準地剔除了次要資訊，讓主旨在摘要中很明主輝，沒有重複的描述，每一句話都承載著必要的資訊量。這種對文字的剖制力顯示出你對主題有著深刻的掌握。閱讀過程非常高效，沒有任何冗餘的幹擾。你對內容的篩選能力非常優秀，值得稱讚。

用字精確：§5| 回饋：你的用字遣詞非常精確，能準確傳達原文的深層意思且無任何模糊之處。像是「血腦屏障」與「中樞神經」等專業詞彙運用得恰到好處，文字表達清晰且符合語境，讓讀者能迅速理解你的意圖。這種對精確度的追求，體現了你嚴謹的學習態度。你的這同造句讓整篇文章顯得專業且具說服力。請繼續保持這種對文字精準度的追求。

修辭：§4| 回饋：你運用了基礎但有效的句式變化，使整體表達具有一定的靈化性。雖然大部分是以直述為主，但文字依然清晰且具備一定的語感。若能嘗試更多同義詞替換或更豐活的轉折，文字會更具魅力。目前的表達已經很得體，能有效強化語意。你已經具備了良好的基礎，只要稍加練習就能更上一層樓。老師相信你一定能寫出更富有文學色彩的文字。

語句流暢性：§5| 回饋：回饋：整體句構自然流暢，標點符號使用正確，語序安排得恰到好處。閱讀時感覺非常順暢，沒有任何生硬或突兀的地方。你對中文語感的掌握非常出色，文字呈現出自然的節奏。這種流暢度讓讀者能專注於內容本身而不會被形式幹擾。你的寫作習慣非常好，展現了良好的語言素養。請繼續保持這種自然且舒適的寫作風格。

改寫：§5| 回饋：你充分展現了語言轉換能力，成功跳脫原文的框架，用自己的話重新詮釋。這不是簡單的抄寫，而是深層理解後的重新建構，非常令人驚艷。你將原文的長句轉化為簡潔的短句，且完整保留了大意。這種**重述**能力證明你已經完全內化了文章的知識。看到你能如此靈活地運用語言，老師感到非常欣慰。這是一個非常高水準的**重述**表現。

圖 3、系統回饋學生科普性文本摘要的內容。

系統回饋內容非常詳盡，此一現象優劣並存，對於高學習動機學生而言，大量回饋可以給予詳細的引導意見，但是對於低學習動機的學生，可能會產生認知負荷，可能導致降低學習意願。關於以上疑慮，且待未來研究進一步調整與驗證。

6 研究限制與未來發展

基於上述發現，開源大型語言模型 Gemma 4 31B 結合評分規準與電腦摘要批改學生摘要，無論在文學性文本或是科普性文本，皆與真人教師的批改結果有高度相關。於教學現場上，教師可以有以下運用：其一，利用系統解決批改過程耗時費力，且無法立即回饋的問題，其二，透過系統協助學生多次練習。

但本研究也存在限制，未來研究方向可從以下方向拓展。其一，擴大實驗對象的年齡層，針對不同學習階段的學生，驗證學生摘要內容於真人教師評分與系統評分的相關性；其二，擴大實驗人數與對象，以更多實驗樣本增加摘要內容多樣性，以驗證系統批改的反應與結果；其三，廣邀更多現場老師，驗證系統使用適用性。

未來研究者可以針對「系統評分回饋」與「學生表現」做進一步研究，學生在系統給與建議之後，是否可以提升摘要分數?是否能在閱讀測驗當中取得更好的表現?同時，系統即時性的回饋是否可以提升學生學習動機?皆是未來研究可以探究的方向。

基於以上結果，得知開源大型語言模型 Gemma 4 31B 結合評分規準與電腦摘要，批改學生於文學性和科普性的長文本摘要寫作所呈現的評分，與真人教師批改結果有高度相關。在即時評分與說明上提供有效回饋，能夠於各個構面上協助學生修改摘要內容，可運用在高中生國語文學習當中長文本摘要、文學與科普等多元文本摘要當中，以期學生在立即性回饋、詳細的回饋指引當中，獲得學習方向，且增進閱讀素養。

本研究承國科會研究計畫 114-2628-H-011 -002 -MY3、國立臺灣師範大學教育部高教深耕計畫華語文與科技研究中心與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補助

謹此致謝

參考文獻

李淑華、洪碧霞（2011）。電腦化課文摘要動態評量系統應用效益之探討。*數位學習科技期刊，3*(1)，86-100。https://www.airitilibrary.com/Article/Detail?DocID=P20170615002-201101-201706220017-201706220017-86-100

林鳳雯、崔夢萍（2011）。網路摘要策略教學與同儕互評對國小六年級學童學習成效之研究。*國民教育，51*(5)，25-34。https://doi.org/10.6476/JNE.201106.0025

馮永敏、賴婷妤與陳美伶 (2015)。國小國語文閱讀摘要教學之實踐與反思。*北市大語文學報*，(4)，61-82。https://www.airitilibrary.com/Article/Detail/P20230705001-N202311030012-00003

黃淇瀅（2020）。*LSA、Word2Vec 與 Google BERT 應用在中文寫作自動評分之成效比較*。〔碩士論文。國立臺中教育大學〕臺灣博碩士論文知識加值系統。https://hdl.handle.net/11296/5u8xgx。

馮樹仁（2001）。*以潛在語意分析法發展摘要寫作評量系統*〔碩士論文，國立臺灣師範大學〕。華藝線上圖書館。https://www.airitilibrary.com/Article/Detail?DocID=U0021-2603200719120095

張國恩、宋曜廷(2005)。*潛在性語意分析及概念構圖在文章摘要和理解評量的應用(3/3)*。國家科學委員會專題計畫成果報告，報告編號 NSC93-2520-S-003-011。台北：行政院國家科學委員會。

藍敏杰（2006）。*潛在語意分析之摘要評量系統—探討特定領域文章、關鍵詞、奇異值保留個數對建構語意空間及評量效度之影響*〔碩士論文，國立臺灣師範大學〕。華藝線上圖書館。https://www.airitilibrary.com/Article/Detail?DocID=U0021-0712200716124994

Abdi, A., Idris, N., Alguliyev, R. M., & Aliguliyev, R. M. (2016). An automated summarization assessment algorithm for identifying summarizing strategies. *PLoS One*, *11*(1), e0145809. https://doi.org/ 10.1371/journal.pone.0145809

Alzamil, L. M., Alhasani, A. M., & Alshehri, S. (2025). Privacy Concerns in ChatGPT Data Collection and Its Impact on Individuals. *Future Internet*, *17*(11), 511. https://doi.org/10.3390/fi17110511

Cezar, B. G. D. S., & Maçada, A. C. G. (2023). Cognitive Overload, Anxiety, Cognitive Fatigue, Avoidance Behavior and Data Literacy in Big Data environments. *Information Processing &*

Management, *60*(6), 103482. https://doi.org/10.1016/j.ipm.2023.103482

Chen, Y., Li, M., Guo, F., & Wang, X. (2023). The effect of short-form video addiction on users' attention. *Behaviour & Information Technology*, *42*(16), 2893–2910. https://doi.org/10.1080/0144929X.2022.2151512

Duke, N. K., & Pearson, P. D. (2009). Effective practices for developing reading comprehension. *Journal of education*, *189*(1-2), 107-122. https://doi.org/10.1177/0022057409189001-208

Gao, M., Hu, X., Yin, X., Ruan, J., Pu, X., & Wan, X. (2025). Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, *51*(2), 661-687. https://doi.org/10.1162/coli_a_00561

Google AI.*Gemma 4 model overview*. GoogleAI. https://ai.google.dev/gemma/docs/core?hl=zh-tw

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Houghton Mifflin. https://digitalcommons.spu.edu/works/173

Huang, Q. Z., Tseng, H. C., & Sung, Y. T. (2025, November). Exploring the Feasibility of Large Language Model-and Rubric-Based Automatic Assessment of Elementary Students' Book Summaries. *In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing* (ROCLING 2025) (pp. 153-166). https://aclanthology.org/2025.rocling-main.19/

Idris, N., Baba, S., & Abdullah, R. (2011). Identifying student's summary writing strategies using summary sentence decomposition algorithm. *Malaysian Journal of Computer Science*, *24*(4), 180–194. https://peradaban.um.edu.my/index.php/MJCS/article/view/6580

Kintsch, E., Steinhart, D., Stahl, G., Lsa Research Group, L. R. 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-B;FT087

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997, December). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. *In Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412-417).

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2), 100017. https://doi.org/10.1016/j.metrad.2023.100017

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

Meta AI. (2024). *Llama 3.1: Open and efficientfoundation language models*. Meta AI.https://huggingface.co/meta-llama

Morris, W., Crossley, S., Holmes, L., Ou, C., Dascalu, M., & McNamara, D. (2025). Formative feedback on student-authored summaries in intelligent textbooks using large language models. *International Journal of Artificial Intelligence in Education*, *35*(3), 1022-1043.https://doi.org/10.1007/s40593-024-00395-0

Nguyen, L., Walters, J., Paul, S., Monreal Ijurco, S., Rainey, G. E., Parekh, N., ... & Darrah, M. (2025). Feeds, feelings, and focus: A systematic review and meta-analysis examining the cognitive and mental health correlates of short-form video use. *Psychological bulletin*, *151*(9), 1125. https://doi.org/10.1037/bul0000498

Ou, L., & Lapata, M. (2025, July). Context-aware hierarchical merging for long document summarization. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 5534-5561).https://aclanthology.org/2025.findings-acl.289/

Ö zdemir, S. (2018). The Effect of Summarization Strategies Teaching on Strategy Usage and Narrative Text 165 Summarization Successi. *Universal Journal of Educational Research 6*(10): 2199-2209.https://doi.org/10.13189/ujer.2018.061018

Pan, G., Chodnekar, V., Roy, A., & Wang, H. (2025). *A Cost-Benefit Analysis of On-Premise Large Language Model Deployment: Breaking Even with Commercial LLM Services (Version 3). arXiv.* https://doi.org/10.48550/ARXIV.2509.18101

Sung, Y. T., Liao, C. N., Chang, T. H., Chen, C. L., & Chang, K. E. (2016). The effect of online summary assessment and feedback system on the summary writing on 6th graders: The LSA‐based technique. *Computers in Education*, *95*, 1 - 18. https://doi.org/10.1016/j.compedu.2015.12.003

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

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

Yao, Q., Zhu, wenwei, Gao, Y., Wang, J., Liu, C., Zhao, G., Wang, Q., 2025. The impact of bullying victimization on short VideoAddiction in adolescents: the role of emotional distressand neural mechanisms. *Addict. Biol*. https://doi.org/10.1111/adb.70038.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, *4*(2), 100211. https://doi.org/10.1016/j.hcc.2024.100211

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, *12*, 39-57. https://doi.org/10.1162/tacl_a_00632

Zhao, W. X., Zhou, K., Li, J., Tang, T., Dong, Z., Hou, Y., ... & Wen, J. R. (2026). A survey of large language models. *Frontiers of Computer Science*, *20*(12), 2012627.https://doi.org/10.1007/s11704-026-60308-3