基於開源大型語言模型設計書籍摘要形成性回饋系統
Designing an Open-Source Large Language Model-Based Formative Feedback System for Book Summaries

詹孟哲 **Meng-Che Chan**
國立臺灣科技大學
數位學習與教育研究所
m11311021@gapps.ntust.edu.tw

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

書籍摘要寫作是一項整合閱讀理解與文本重構的語文任務。然而，新手寫作者常出現內容誤解、重點遺漏或修訂停留於表層文字修改等問題。雖然近年自動摘要回饋技術逐漸發展，但現有系統較少兼顧修訂順序規劃與形成性引導，難以依合理順序推進學習者進行摘要修訂。爲此，本研究設計一套基於開源大型語言模型（Gemma 4 31B）之書籍摘要形成性回饋系統。系統將摘要品質問題區分爲誤解、邏輯、遺漏、主觀、多餘及破碎鬆散六類，並依據由內容理解至語篇表達的修訂原則安排回饋順序。每輪回饋聚焦於一項當前最優先處理的問題，透過提示、線索、提問與選擇題等鷹架策略，引導學習者自行發現與修正問題，而非直接提供標準答案。本研究採專家評估方式進行初步驗證，邀請一位在職國文教師依八項教學能力評估構面，進行 157 筆逐輪回饋評估。結果顯示，在三點量表下，各項指標平均分數皆達 2.91 分以上，其中，系統在最終答案揭露構面中能避免直接提供完整答案，且連貫性、教練語氣及自然度等指標皆獲滿分評價。評估結果也顯示，系統大致能依預定的問題診斷與回饋原則運作。研究結果初步支持大型語言模型應用於書籍摘要形成性回饋與數位閱讀寫作教練之可行性。

Abstract

Book summary writing requires learners to integrate reading comprehension with text reconstruction. However, novice writers often produce summaries containing misunderstandings and omissions, and their revisions may remain limited to surface-level textual changes. Although automated feedback technologies for summary writing have advanced in recent years, existing systems seldom consider both revision sequencing and formative guidance, making it difficult to support learners in revising summaries in a pedagogically appropriate order. To address these issues, this study proposes a formative feedback system for book summary writing based on the open-source large language model Gemma 4 31B. The system categorizes summary quality issues into six types: misunderstanding, logical inconsistency, omission, subjectivity, redundancy, and incoherence. Following a revision sequence from content comprehension to discourse expression, each feedback round focuses on a single priority issue and employs scaffolding strategies such as prompts, hints, questions, and multiple-choice guidance to support learner revision without directly revealing answers. A preliminary expert evaluation was conducted using 157 iterative feedback records. One in-service Chinese language teacher assessed the system based on eight pedagogical dimensions. Results showed that mean scores across the eight dimensions ranged from 2.91 to 3.00 on a three-point scale. The system received perfect scores for avoiding direct answer revelation under the Revealing of the Answer dimension, as well as for Coherence, Tutor Tone, and Human-likeness. The findings provide preliminary evidence supporting the feasibility of applying large language models to formative feedback for book summary revision.

關鍵字：書籍摘要、形成性回饋、自動化寫作回饋、大型語言模型、閱讀寫作教學

Keywords: Book Summary Writing, Formative Feedback, Automated Writing Feedback, Large Language Models, Reading and Writing Instruction

1 緒論

摘要寫作（summary writing）被視爲閱讀與寫作整合的重要語文任務。學習者需在理解來源文本後，辨識重要概念，篩選關鍵資訊，並以自己的語言重新組織內容。從認知訊息

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 322–332
November 6-7, 2026 ©2026 Association for Computational Linguistics

處理的觀點來看，摘要寫作是檢驗閱讀理解與文本重構能力的重要認知過程 (Kintsch and Van Dijk, 1978)。然而，摘要寫作對新手寫作者而言具有相當的挑戰性。研究指出，第二語言學習者在進行摘要時，常出現無法掌握重點、遺漏重要資訊、過度依賴原文措辭，甚至因理解錯誤而產生內容偏差與事實扭曲等問題 (Chuenchaichon, 2022)。

除了作爲寫作任務之外，摘要也常被應用於閱讀理解評量。由於摘要品質能反映學習者對文本內容的理解程度及資訊整合能力，過去研究曾利用摘要寫作任務建立自動評分模型，以衡量閱讀理解表現 (Madnani et al., 2013)。因此，當學習者摘要出現內容偏差、重點缺漏或組織混亂等問題時，往往不僅代表寫作表達上的困難，也可能反映其對來源文本的理解存在偏差。

摘要寫作品質的提升是一個需要不斷進行認知思考與反覆修訂的歷程。在寫作修訂的研究中，Sommers (1980) 指出新手寫作者與成熟寫作者在修訂策略上存在顯著差異。新手傾向於將修訂局限於字詞替換與表層語句的潤飾，而成熟寫作者則將修訂視爲遞迴思考的過程，關注文章的整體意義、論點發展與結構重組 (Sommers, 1980)。將此寫作理論延伸至摘要範疇，意味著高品質的摘要修訂除了字句修飾，也涉及內容理解與結構層面的調整。此外，Hattie and Timperley (2007) 提出的形成性回饋模型指出，有效回饋的核心在於協助學習者回答「目標爲何（Where am I going?）」、「目前進展爲何（How am I going?）」以及「後續如何改善（Where to next?）」三個核心問題，進而縮短當前表現與目標之間的差距 (Hattie and Timperley, 2007)。因此，相較於僅給予分數或等第的總結性評定，能夠引導學習者辨識問題並支持修訂決策的形成性回饋，更能顯著促進學習歷程。

隨著自然語言處理與大型語言模型（Large Language Model, LLM）技術的演進，自動化寫作評估（Automated Writing Evaluation, AWE）系統的教學定位正發生變化。過去傳統 AWE 系統多著重於可量化的表層語言特徵與公式化寫作，能提供即時回饋並協助學生修正表層錯誤，也具有支援教師課堂管理的潛力。然而，這類系統較難評估寫作的美學特質、內容知識及深層意涵，長期使用亦可能使學生的寫作趨向機械化與公式化 (Ware, 2011)。爲突破傳統 AWE 的局限，現代教學系統轉而探討如何運用具備強大語境理解能力的 LLM，提供更具深層引導性的形成性回饋。

然而，自動化回饋的教學效益取決於其訊息遞送策略。依據 Shute (2007) 對形成性回饋的探討，回饋應扮演「鷹架（Scaffolding）」角色，透過提示、提問與線索等促進式回饋（Facilitative feedback），引導學習者自行建構理解與修正思考 (Shute, 2007)。此外，根據 Sweller (1988) 提出的認知負荷理論（Cognitive Load Theory），人類的工作記憶容量極爲有限，當任務資訊過多或處理歷程過於複雜時，過高的認知負荷將阻礙基模建構（Schema acquisition）與有效學習 (Sweller, 1988)。因此，在自動化回饋設計中，若一次提供過多修正資訊極易引發認知超載。本研究採單一問題聚焦策略，藉由控制單次回饋資訊量並建立問題處理的優先順序，協助學習者降低認知負擔並逐步完成修訂。

綜上所述，本研究設計一套以開源大型語言模型爲核心的書籍摘要形成性回饋系統。該系統將學習者摘要中的常見問題歸納爲六種類型，並依據修訂重要性，建立由「內容理解」、「結構精煉」至「語篇表達」的修訂順序。在互動機制上，系統採取多輪形成性回饋策略，每輪僅聚焦於當前最需優先處理的一項問題，透過提示、線索、引導性提問與選擇題等鷹架手段，支持學習者自主進行摘要修訂。

本研究旨在檢驗此一結合「摘要問題分類」、「修訂優先順序」與「形成性鷹架策略」之回饋架構，是否能有效約束大型語言模型的回饋行爲，提供具備教學價值與認知適性之摘要修訂支持。

2 文獻探討

2.1 摘要自動批改系統與回饋設計現況

摘要寫作兼具閱讀理解與文本重構特性，因此除作爲寫作任務之外，也常被應用於閱讀理解評量。相較於一般客觀測驗，摘要評量能更直接反映學習者是否掌握文本核心概念，以及是否能將重要資訊重新組織爲完整內容。然而，摘要評量需逐篇檢查內容正確性、重點涵蓋情形及整體品質，因此人工評量往往耗費大量時間與人力資源。爲提升評量效率，研究者逐漸運用自然語言處理技術發展摘要自動評量系統 (Madnani et al., 2013)。

在國際研究方面，早期摘要自動評量系統多採語意相似度分析技術。其中，Wade-Stein and Kintsch (2004) 所開發的 Summary Street 利用潛在語意分析（Latent Semantic Analysis, LSA）比較學習者摘要與來源文本之間的高維度語意向量相似度，並針對內容涵蓋、相關性及冗餘資訊提供即時回饋。然而，此類系統主要依賴統計語意向量表徵，未進一步解析句

法結構與語序資訊，且早期多側重於評量與內容涵蓋度追蹤。隨著研究焦點逐漸由評量轉向學習支持，形成性回饋的重要性逐漸受到重視。Hattie and Timperley (2007) 指出，回饋的重要價值不在於呈現學習結果，而在於協助學習者理解目前表現與學習目標之間的差距。Shute (2007) 進一步指出，有效的形成性回饋應具備具體性、任務導向及可操作性，避免直接提供完整答案，以促進學習者主動建構知識。寫作教育領域針對中小學階段的後設分析也顯示，無論來自教師、自我評量、同儕或電腦系統的回饋，皆能顯著提升學習者寫作品質 (Graham et al., 2015)。形成性回饋觀點也逐漸應用於摘要修訂情境。Barkela and Leuchter (2024) 所提出的 FALB（Formative Automated Learning for Summarizing）系統，能針對內容涵蓋、相關性、篇幅控制、避免冗餘以及直接複製原文等面向提供形成性回饋。研究結果顯示，自動形成性回饋能有效促進學習者投入摘要修訂歷程。特別在引導學習者符合篇幅限制與降低直接複製原文上效果顯著，但在改善「避免冗餘」上的提升則相對有限，主因是系統缺乏寫作前企劃階段（planning phase）的提示引導。Sawaki et al. (2024) 所發展的線上摘要形成性評量模組，透過重要資訊標示、大綱對比、規則自評與逐字複製檢測等方式提供即時內容回饋。研究結果顯示，此類回饋能促使學習者進行具意義的修訂活動，且學習者與教師皆高度肯定其教學價值。

在臺灣研究方面，摘要回饋系統的發展也呈現由評量走向形成性回饋的趨勢。陳佳琳 (2013) 以潛在語意分析技術結合專家概念圖及概念詞回饋機制，發展線上摘要評量與回饋系統。研究結果顯示，就單一回饋模式而言，語意回饋與概念詞回饋皆能顯著提升學習者摘要表現，且概念層次與語意指引的回饋比單純分數更具有學習價值。然而，研究亦發現同時提供「語意與概念詞」之雙重回饋未產生顯著加乘效果，顯示系統設計需考量資訊密度與初學者的認知負擔。研究也指出，摘要系統應兼顧評量結果與回饋設計，以支持學習者後續修訂歷程。此外，也有形成性評量研究證實回饋對學習歷程的重要性。李貞誼 (2021) 將形成性評量回饋策略應用於國中閱讀教學，研究結果顯示形成性回饋能有效提升學習者自我調整學習能力。雖然在量化「閱讀能力測驗」上兩組未達顯著差異，但質性訪談與問卷分析顯示其有助於提升學習者閱讀自信心。同時，學習者對回饋教學持正面態度。此結果顯示，回饋不僅具有診斷功能，若能結合具體的策略引導，更能作爲支持學習者持續學習與自我監控策略的重要工具。

近年大型語言模型的發展進一步改變摘要評量與回饋的研究方向。Botarleanu et al. (2022) 證實 Transformer 類模型已具備良好的摘要多任務自動評分能力。在寫作回饋的應用上，Meyer et al. (2024) 於議論文寫作情境中發現，大型語言模型產生的回饋能提升學習者的修訂表現、寫作動機及正向情緒。Morris et al. (2025) 則進一步將分析式評分規準降維爲內容與措辭構面，並利用大型語言模型進行預測，整合至智慧教科書中提供摘要形成性回饋。國內研究方面，Huang et al. (2025) 指出大型語言模型已能提供具體且即時的書籍摘要回饋，但在高層次語言表達與重述能力評估上仍存在限制。

綜上所述，摘要自動批改系統已逐漸由著重評量與評分功能的工具，發展爲支持學習者修訂歷程的形成性回饋系統。而大型語言模型的發展則進一步提升了系統在文本理解與自然語言生成方面的能力。然而，現有研究較少探討當學習者摘要同時存在多項問題時，系統應如何決定修訂優先順序、控制單輪回饋資訊量，以及透過何種鷹架策略引導學習者自主完成修訂。因此，本研究將進一步探討摘要品質問題分類、修訂順序與形成性鷹架策略，作爲後續系統設計的理論基礎。

2.2 摘要品質問題分類與修訂順序

除了回饋內容本身之外，系統如何定義摘要品質問題，以及當摘要同時存在多項問題時，應優先處理何種問題，也是影響修訂成效的重要因素。

摘要評估不同於一般寫作評量。評量者除了關注語言品質之外，也需檢查摘要是否忠實反映來源文本、是否涵蓋重要資訊，以及是否維持適當的資訊組織與篇幅控制。Yamanishi et al. (2019) 指出，摘要評量需同時考量內容、改寫技巧（Paraphrasing）與語言使用等多個維度。Sladoljev-Agejev and Šnajder (2017) 也強調，摘要寫作兼具閱讀理解（Reading-for-Understanding）與寫作產出（Reading-to-Write）雙重構面，採用分析式評量架構能更細緻地涵蓋內容正確性與語篇組織品質。

近年來，自然語言處理領域逐漸由整體評分轉向細粒度摘要分析。在摘要品質研究中，忠實性（Faithfulness）被視爲最基本的品質要求。Maynez et al. (2020) 指出，摘要即使具備流暢且合理的表達，仍可能產生與來源文本不一致的內容。Pagnoni et al. (2021) 進一步發現，摘要中的事實錯誤可能涉及人物、事件、時間或事件關係等不同層面的偏差。另一方面，Steen and Markert (2022) 指出，摘要

是否具備良好的語篇連貫性（Coherence），也是影響摘要品質的重要因素。

此外，Song et al. (2024) 提出的 FineSurE 架構將摘要品質區分爲忠實性（Faithfulness）、完整性（Completeness）與簡潔性（Conciseness）等主要面向。統整摘要品質評估、摘要事實性研究、摘要連貫性研究以及摘要評分規準相關文獻，本研究進一步將文獻中的品質構面轉爲形成性回饋情境之六類摘要品質問題，並依其影響程度分爲內容理解、結構與精煉，以及語篇表達三個修訂層次，如表 1 所示。

修訂層次	品質面向	系統問題類型	對應概念
內容理解	忠實性	誤解	事實理解錯誤
	忠實性	邏輯	關係錯置
結構與精煉	完整性	遺漏	重要資訊缺失
	摘要規範	主觀	客觀性不足
	簡潔性	多餘	冗餘資訊
語篇表達	連貫性	破碎鬆散	連貫不足

表 1: 摘要品質面向、本系統定義的問題類型與修訂層次之對應關係

如表 1 所示，本研究將摘要品質問題區分爲三個修訂層次。

在內容理解層次，本研究根據忠實性（Faithfulness）相關研究區分爲「誤解」與「邏輯」兩類問題。誤解對應於事實理解錯誤（factual misunderstanding），指學習者對來源文本的重要人物、事件、概念或事實產生錯誤理解；邏輯則對應於關係錯置（relation inconsistency），指因果、條件、時間順序或事件關係出現偏差 (Maynez et al., 2020; Pagnoni et al., 2021)。由於此類問題直接影響摘要是否正確反映原文內容，因此被歸類爲最優先處理的修訂層次。

在結構與精煉層次，本研究結合完整性（Completeness）、簡潔性（Conciseness）與摘要規範等面向。其中，「遺漏」對應於重要資訊缺失（information omission），指摘要缺少重要訊息、核心概念或關鍵事件，以致文本內容呈現不完整 (Song et al., 2024)；「主觀」對應於客觀性不足（lack of objectivity），指摘要混入個人感受、評論或價值判斷，而未維持摘要應有之客觀呈現特性 (Morris et al., 2025)；「多餘」則對應於冗餘資訊（redundancy），指摘要保留過多與核心主旨關聯較低的細節，降低資訊濃縮效果 (Song et al., 2024)。

在語篇表達層次，本研究結合語篇連貫性（Coherence）與銜接性（Cohesion）相關研究定義「破碎鬆散」問題，其對應於連貫與銜接不足（coherence and cohesion deficiency），指句子之間缺乏適當銜接標記或概念重疊、資訊排列不合理，影響整體語篇理解 (Steen and Markert, 2022; Burkhart et al., 2020)。雖然此類問題會影響摘要可讀性，但通常建立在內容已大致正確的前提下，因此安排於較後續的修訂階段。

綜上所述，本研究將文獻中的摘要品質構面轉化爲誤解、邏輯、遺漏、主觀、多餘及破碎鬆散六類系統問題，並進一步劃分爲內容理解、結構與精煉，以及語篇表達三個修訂層次，作爲後續系統回饋設計的基礎。

然而，形成性回饋除了辨識問題之外，需要決定當摘要同時存在多項問題時，應優先處理何種問題。如前述 Sommers (1980) 發現，新手寫作者傾向於將修訂視爲字詞替換或局部修改，而成熟寫作者則會優先處理內容意義、文章組織與整體架構，待核心內容確立後，才進一步調整語言表達。另外，Faigley and Witte (1981) 將修訂區分爲「表面修訂（Surface Changes）」與「文本基礎修訂（Text-Based Changes）」，後者包含對文章概念理解與宏觀結構（Macrostructure）的調整。文本基礎與宏觀層面的修訂主要涉及內容增刪、概念理解與組織結構調整；表面修訂則著重於句型、措辭及語法修飾。此觀點顯示，寫作修訂可視爲由宏觀至微觀（macro-to-micro）的歷程，即先確認文本是否正確表達核心內容，再逐步處理資訊組織與語篇表達問題。

近年研究也指出，當學習者修訂任務過於複雜時，若能將修訂目標聚焦於單一且最重要的問題，較能促進有效修訂。Jansen et al. (2024) 發現，相較於僅提供分數或評分規準，依據學習者表現設定個人化修訂目標，更能提升學習者後續修訂品質。此結果顯示，建立合理的修訂優先順序有助於引導學習者聚焦於當前最重要的修訂任務。

因此，本研究結合摘要品質面向與修訂理論，採用由宏觀至微觀（macro-to-micro）的修訂原則。當摘要同時存在多項問題時，系統優先處理對摘要核心意義影響較大的內容理解問題，再逐步推進至結構與精煉以及語篇表達層面，以支持學習者循序漸進地完成摘要修訂。

2.3 認知負荷與形成性鷹架策略

雖然決定修訂順序有助於建立修訂焦點，但形成性回饋的有效性仍取決於學習者是否能理解與運用回饋資訊。Shute (2007) 指出，有效形成性回饋應具體、任務導向且可操作，同時避免直接提供完整答案，以促進學習者主動參與思考與問題解決。

然而，在摘要修訂情境中，學習者需同時面對內容理解、資訊選擇與語言表達等多重挑戰。根據認知負荷理論，當學習者於同一時間需處理過多資訊時，有限的工作記憶資源將受到壓迫，進而影響學習成效 (Sweller, 1988)。摘要回饋系統的研究回顧亦指出，早期摘要輔助系統（如 Summary Street）若同時呈現過多類別的回饋資訊，容易使學習者產生資訊過載（information overload），反而削弱回饋的實際效益 (Sung et al., 2016)。

有鑑於認知負荷理論所強調之有限工作記憶觀點，以及形成性回饋應以「可管理單元（manageable units）」呈現之原則 (Shute, 2007)，本研究推導並採取單一問題聚焦策略。每輪回饋針對當前最優先處理的一項問題進行引導，而非一次呈現所有修訂要求，以降低學習者同時處理多重修改目標時的認知負荷。

另一方面，鷹架式引導強調透過適度提示協助學習者完成原本無法獨立完成的任務。Chen et al. (2014) 在學習引導研究中提出漸進式提示（Progressive Prompting）策略，透過分階段提示逐步引導學習者，研究顯示漸進式提示能有效提升學習成就並促使學習者投入更多心力進行深度思考。雖然該研究原應用於自然科觀察任務，但其「當學習者未達目標時，再逐步提供更具體引導」的核心理念，同樣適用於文字摘要修訂，能兼顧學習支持與自主修訂之間的平衡。

綜上所述，本研究根據不同摘要問題設計相應的鷹架策略：針對誤解與邏輯的問題，提供較直接的選項提示；針對遺漏的問題，提供內容線索區塊；針對主觀與多餘的問題，透過提問與比較，引導學習者判斷資訊重要性；針對破碎鬆散的問題，則提供語篇組織與銜接連接詞提示範例。本研究整合摘要品質問題分類、由宏觀至微觀的修訂順序、認知負荷控制與漸進式鷹架理念，建立大型語言模型書籍摘要形成性回饋系統的理論架構，以支持學習者逐步完成摘要修訂。

3 實驗方法

本研究採用系統建置與專家評估的研究設計，旨在檢驗基於開源大型語言模型的書籍摘要回饋系統是否能提供具診斷性與教學可用性的形成性回饋。本研究主要檢視四項問題：第一，系統是否能正確辨識學習者摘要中的主要問題；第二，系統是否能依合理修訂順序，每輪聚焦於單一問題提供回饋；第三，系統是否能在不直接揭露答案的情況下，提供具體且可操作的有效修訂引導；第四，在多輪修訂歷程中，前一輪所聚焦的目標問題是否於下一輪摘要中獲得處理，以及系統是否能在該問題獲得修正後推進至後續待修訂問題。

本系統定位為閱讀寫作教練，其主要功能並非直接替學習者撰寫摘要，而是協助學習者逐步理解並處理摘要中的問題。由於本研究採專家模擬修訂，而非真實學習者教學實驗，因此上述多輪歷程分析主要檢視目標問題是否獲得處理，不直接推論學習者摘要或閱讀理解能力是否提升。

3.1 系統設計

3.1.1 系統運作流程

本系統以 Gemma 4 31B 開源大型語言模型為基底於本機部署，使用預先設定的系統提示詞運作，作為書籍摘要形成性回饋教練。學習者登入系統後選擇已閱讀的書籍，送出撰寫完成的摘要後，取得系統回饋。若系統判定摘要存在待修訂的品質問題，則學習者依據回饋內容進行修改並重新提交。若摘要已符合要求，系統即結束修訂流程並提供最終評分。整體系統運作流程如圖 1 所示。

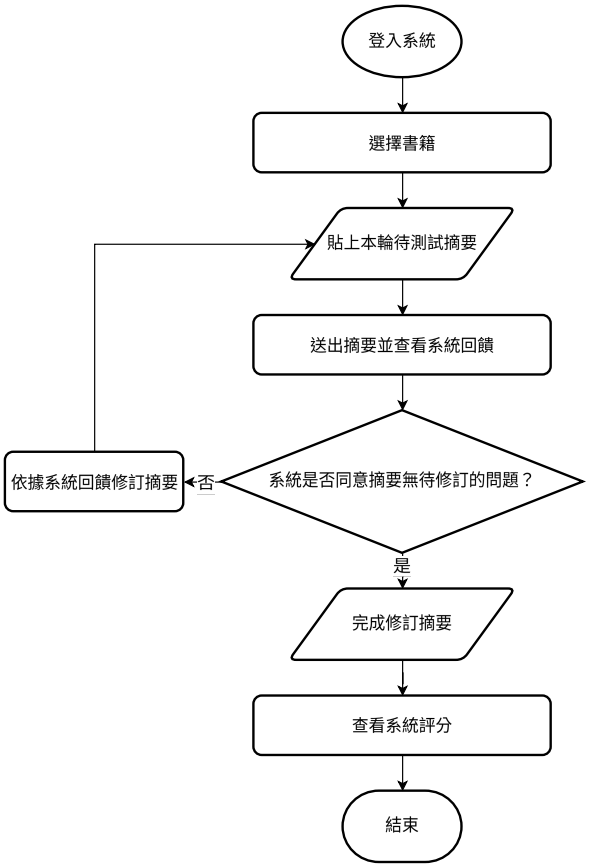

圖 1: 系統運作流程圖

3.1.2 摘要品質診斷與修訂順序

依據前一章節所述之修訂理論，本研究將摘要品質問題區分為六類，包括：(1) 誤解、(2)

邏輯、(3) 遺漏、(4) 主觀、(5) 多餘，以及 (6) 破碎鬆散。

系統讀取書籍原文後，自動生成該書籍的關鍵事實、理想摘要範文與條列式重點，作爲判斷依據。依循「先處理內容理解，再處理結構與精煉，最後處理語篇表達」的原則，逐項排查六類摘要品質問題。六類問題進一步歸納爲三個修訂層次。第一層次爲內容理解，包含誤解與邏輯問題；第二層次爲結構與精煉，包含遺漏、主觀與多餘問題；第三層次爲語篇表達，包含破碎鬆散問題。當摘要同時存在多項問題時，系統主要選擇處理優先層級最高的一個摘要問題，並透過相應的鷹架策略引導修訂，以避免一次呈現過多修改要求而增加學習者的認知負荷。摘要經修改並重新提交後，系統再次診斷修訂後文本，檢查前一輪所聚焦的目標問題是否仍然存在。若該問題已獲得修正，系統再依修訂順序處理下一項待修訂問題。若問題仍然存在，則繼續針對該問題提供引導。

系統透過「診斷優先問題、提供單一問題回饋、重新檢查修訂文本，以及推進至下一問題」的循環，依序處理摘要中的待修訂問題。此循環著重於目標問題是否獲得處理。當系統判定摘要已無需優先處理的問題時，即停止要求修改並給予肯定。

3.1.3 形成性回饋與鷹架設計

系統依據問題類型提供相應的形成性回饋與鷹架引導。三個修訂層次及其對應策略如表 2 所示。內容理解問題透過指出矛盾與選擇題協助學習者重新判斷；遺漏問題提供內容線索引導回想原文；主觀與多餘問題透過提問協助辨識核心與次要資訊；破碎鬆散問題則提供語篇組織與銜接連接詞等提示。透過上述設計，系統引導學習者自主修訂，而非直接提供答案。

修訂層次	問題類型	引導方式
內容理解	誤解、邏輯	指出矛盾或關係錯誤，提供選擇題引導學習者判斷。
結構與精煉	遺漏	提供內容線索，引導學習者回想缺漏重點。
	主觀、多餘	引導區辨核心資訊與次要資訊，思考是否保留。
語篇表達	破碎鬆散	提供句型或連接詞提示，協助重組語句。

表 2: 修訂層次、問題類型與引導方式

3.1.4 系統輸出設計

系統每輪輸出包含兩個部分。第一部分爲「結構化診斷紀錄」，記錄系統依序檢查六類摘要品質問題後的判斷結果、最終選定之摘要問題類型，以及所採用的引導策略。第二部分爲「學習者可見鷹架回饋」，包含具體讚美、提示、線索與修改方向，用以引導學習者自主修訂摘要。

在實際教學情境中，系統僅呈現第二部分內容，學習者不會看見系統內部診斷過程。此外，爲利本研究的專家檢驗系統診斷的結果與回饋內容是否一致，在專家評估模式中會同時呈現上述兩個部分，使專家得以檢視系統如何由診斷結果產生相應之形成性回饋。其運作範例如表 3 所示。

3.2 專家評估設計

本系統之最終應用對象爲學習者，然而，本研究階段目的在於驗證形成性回饋機制是否能依預定教學原則穩定運作，而非直接檢驗學習者學習成效。因此，本研究定位爲系統設計之初步驗證。在正式導入學習者使用之前，先由具相關經驗之教師進行專家評估，檢視問題辨識、修訂順序與回饋引導之合理性。建立系統回饋品質之初步根據後，未來再以學習者爲對象進行教學實驗，以檢驗系統對摘要修訂歷程與學習成效之影響。

3.2.1 評估材料書籍

本研究選取兩本主題不同的中文書籍作爲評估材料，分別涵蓋社會議題以及文學與科學議題，以檢驗系統於不同內容類型中的摘要回饋表現。

3.2.2 專家背景

本研究邀請一位在職國文教師擔任專家評估者。該教師具備中文閱讀與寫作教學經驗，能依據摘要文體要求與學習者語文程度，判斷系統指出的摘要問題是否合理，以及回饋內容是否能有效協助學習者進行修訂。

3.2.3 評估樣本設計

專家針對書籍設計 9 類標準化摘要測試樣本。樣本類型涵蓋六類單一問題、兩類複數問題，以及一類無問題摘要，用以檢驗系統在不同摘要狀態下的診斷穩定性、複數問題情境中的處理順序，以及摘要完成狀態的判定能力。樣本設計如表 4 所示。

針對 9 類標準化摘要測試樣本，每一個樣本會測試 2 次，因此每一本書籍測試共 18 次。兩本書籍的評估共包含 36 次獨立互動任務。

學習者輸入摘要（節錄）
李白非常喜歡喝酒，常常在喝酒時想像明月和自己的影子陪伴自己。作者提到，酒精會讓人變得更聰明，所以李白才能寫出許多著名的詩作。我覺得李白眞的很厲害，喝酒竟然還能幫助創作。
系統結構化診斷紀錄（評估模式可見）
系統判定此摘要同時包含「誤解」與「主觀」等問題。其中，「酒精會讓人變得更聰明，所以李白才能寫出許多著名的詩作」與來源文本內容不符。原文指出酒精會影響大腦運作，使人產生愉悅、自信與奔放的情緒，而非提升智力。依據由宏觀至微觀與單一問題聚焦原則，系統本輪選擇優先級較高的「1. 誤解」作爲最終診斷，暫不要求學習者同時處理主觀評論問題；採用策略爲顯性指令鷹架。
學習者可見鷹架回饋（節錄）
哈囉！你有抓到李白喜歡飲酒，以及他會透過想像明月和影子陪伴自己，這部分做得很好喔。不過，我們一起再想想看：文章提到酒精對李白產生了什麼影響呢？
(A) 酒精提升了李白的智力，使其更能創作詩歌。
(B) 酒精主要影響李白的情緒與行爲表現，而非提升智力。
請先把這個最重要的內容修改正確，再把完整摘要重新傳給我看一次喔！

表 3: 系統診斷與鷹架回饋範例

樣本編號	樣本類型	檢驗目標與預設問題
S01	單一問題	檢驗「誤解」
S02	單一問題	檢驗「邏輯」
S03	單一問題	檢驗「遺漏」
S04	單一問題	檢驗「主觀」
S05	單一問題	檢驗「多餘」
S06	單一問題	檢驗「破碎鬆散」
S07	複數問題 A	檢驗處理順序：誤解 > 遺漏
S08	複數問題 B	檢驗處理順序：邏輯 > 主觀
S09	無問題摘要	檢驗完成狀態的判定

表 4: 評估樣本設計與檢驗目標

每一筆測試任務由一篇專家設計的初始摘要測試樣本開始，專家依據系統回饋模擬學習者逐輪修訂摘要。每送出一次摘要，即形成一筆回饋紀錄。當系統判定摘要完成且專家同意，或互動達到 8 輪測試上限時，即完成該筆摘要任務。本研究共包含 36 個獨立互動任務，並由這些任務產生 157 筆逐輪回饋紀錄。逐輪回饋紀錄主要用於分析單輪回饋品質。完整辨識、修訂順序、問題推進與完成判定則以 36 個互動任務爲主要分析單位。由於同一互動任務中的多輪紀錄具有相依性，157 筆紀錄不視爲彼此獨立的實驗樣本。

3.2.4 資料蒐集流程

本研究透過 Google 表單蒐集資料。專家先設計摘要樣本並送入系統取得回饋，每輪互動後記錄回饋品質，當系統判定完成，且專家也同意，或是達八輪測試上限時結束測試，再進行整體歷程評估。整體流程如圖 2 所示。

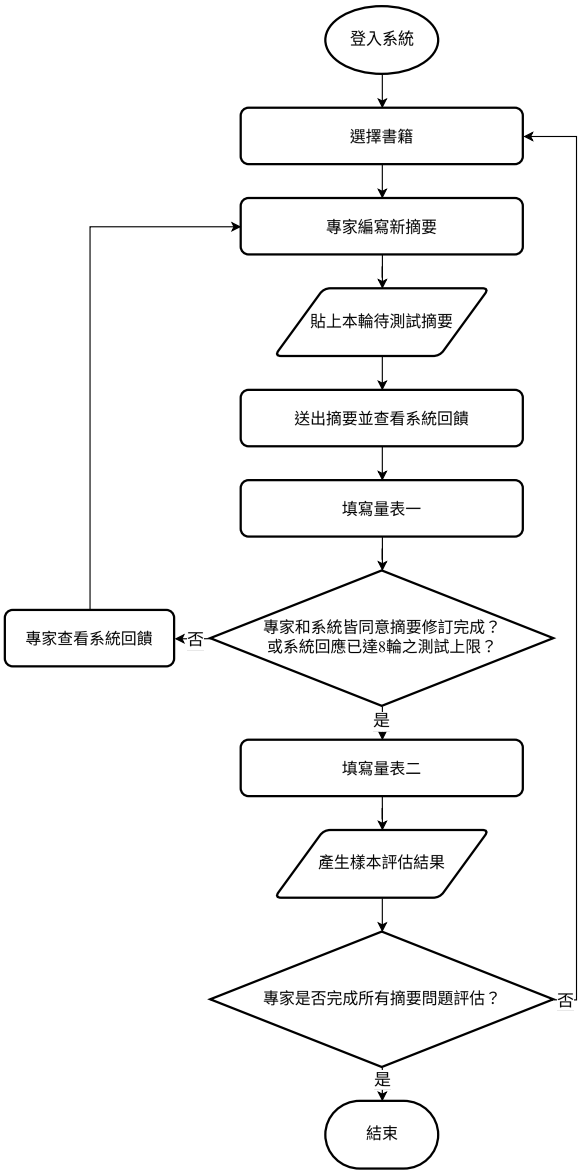

圖 2: 專家評估流程圖

3.2.5 評估工具與構面設計依據

本研究之專家評估工具以 Google 表單建置，包含量表一與量表二。量表一用於評估系統逐輪回饋之表現，由專家於每次送出摘要並取得系統回饋後填答。評估內容包括該輪摘要文本、專家認爲應優先處理之問題、系統實際處理之問題、系統是否聚焦於單一問題，以及系統所指出之問題是否確實存在於摘要中。摘要問題分爲誤解、邏輯、遺漏、主觀、多餘及破碎鬆散等六類，用以比較專家判斷與系統實際處理之問題。

量表一另設有回饋情況檢核題，供專家判斷系統回饋是否出現問題辨識不完整、同輪處理多個問題、直接提供完整摘要、問題定位不明確、缺乏具體修改方向、語氣不當或表達不自然等情況，並評定其涉及之回饋品質構面。此外，量表一亦包含系統回覆對目標學生之協助程度、該輪評語，以及是否繼續進入下一輪等題項。

量表二於單一摘要樣本之多輪評估完成後填寫，用以記錄系統回饋輪數、已辨識及未辨識之問題類型，並評估系統於整體互動歷程中的問題辨識完整性、單一問題聚焦程度、問題處理順序合理性、修訂推進能力、完成判定合理性、對目標學生之協助程度及整體可接受性。其中，完成判定合理性係指完成應處理之摘要問題後，系統是否適時停止要求修改並給予肯定。

量表一之逐輪回饋品質構面係參考 Maurya et al. (2025) 提出之 AI 家教教學能力評估分類學（Evaluation Taxonomy），並依摘要修訂情境調整其操作型定義。八項構面包括：（1）問題辨識（Mistake Identification）、（2）問題定位（Mistake Location）、（3）最終答案揭露（Revealing of the Answer）、（4）提供引導（Providing Guidance）、（5）可操作性（Actionability）、（6）連貫性（Coherence）、（7）教練語氣（Tutor Tone），以及（8）自然度（Human-likeness）。其中，問題辨識與問題定位係指系統能否辨識摘要中待修訂之問題並指出其位置。最終答案揭露則用以評估系統是否直接提供完整摘要，或提供足以取代學生自行修訂之答案。

上述分類學係依據主動學習、學習者目標與需求、認知負荷與後設認知，以及學習動機等學習科學原則建構，並透過錯誤辨識與定位、避免過早揭露答案、提供適當引導、指出後續行動及採取鼓勵性語氣等可觀察之回應特徵進行評估 (Maurya et al., 2025)。原分類學主要應用於數學教育對話中的錯誤修正任務，本研究則依摘要修訂任務與系統設計目標，調整各構面之操作型定義及評分標準。

在 Maurya et al. (2025) 之分類學試行驗證階段，四位標註者檢視分類架構與標註指引，未提出需增列或刪除之構面。該階段之 Fleiss' kappa 爲 0.65，後續正式標註之平均 Cohen's kappa 爲 0.71；依原研究採用之判準，兩者皆達 substantial agreement，顯示該分類學於原研究情境中具有一定程度之評分者間一致性。

八項逐輪回饋品質構面，以及量表一與量表二中評估系統表現之整體判斷題，皆採三點量尺評定。爲使各題項之評分方向一致，本研究將 3 分設定爲符合預期標準、2 分爲部分符合、1 分爲不符合，並依各構面之操作型定義訂定評分判準。就「最終答案揭露」構面而言，分數越高表示系統越能避免直接提供可取代學生自行修訂之答案。其餘有關問題類型、回饋情況及互動歷程之題項，則依評估目的採單選、複選、流程判斷或開放式作答。

4 實驗結果

4.1 回饋品質與形成性指標表現

表 5 呈現 157 輪回饋在八項教學能力評估構面上的專家評分結果。數據顯示，系統在形成性回饋的各項指標上表現穩定。特別在「最終答案揭露」（Revealing of the Answer）構面中，系統避免直接揭露答案的平均得分爲 3.00；「連貫性」、「教練語氣」與「自然度」等構面的平均分數亦皆爲 3.00。此結果顯示系統能在避免直接代寫的情況下提供自然且具教學性的修訂引導。

八項教學能力評估構面	平均分數 (滿分 **3**，越高越好)
問題辨識 (Mistake Identification)	2.91
問題定位 (Mistake Location)	3.00
最終答案揭露 (Revealing of the Answer)	3.00
提供引導 (Providing Guidance)	2.97
可操作性 (Actionability)	2.97
連貫性 (Coherence)	3.00
教練語氣 (Tutor Tone)	3.00
自然度 (Human-likeness)	3.00

表 5: 逐輪系統回饋品質專家評分結果

4.2 系統診斷表現與侷限分析

在 157 筆逐輪紀錄中，專家評估結果顯示系統大致能依預定機制運作。如表 6 所示，共有 127 筆紀錄被判定爲無明顯缺失。在多數情況下，系統能聚焦於單一核心問題提供回饋，僅有 7 筆出現同一輪要求學習者修改多個問題的情形。此外，對於誤解、遺漏等較明確且具體的摘要問題，系統多能正確辨識並提供相應引導。針對複數問題樣本，系統也大致能依預定順序逐步推進修訂歷程。專家評估指出，系統回饋語氣自然且具有教練特性，能以學習者較容易理解的方式說明問題並鼓勵持續修訂。

然而，專家評估也指出系統在特定條件下仍存在侷限。由表 6 可知，較常發生的情況包括未辨識出本輪合理或主要的摘要問題（14 筆）以及摘要已接近完成，但系統仍要求修改（8 筆）。進一步檢視案例後發現，當摘要同時包含多項問題時，系統有時傾向聚焦於較明顯的問題，相較不明顯的問題類型可能被忽略。此外，主觀評論、過度詮釋、多餘資訊及部分邏輯問題等語意邊界較爲模糊的情境，系統較容易出現診斷不一致或判斷偏嚴格的情形，導致完成狀態判定不夠穩定。部分回饋雖能指出問題，但在語句銜接或修訂策略等面向的操作性引導仍有改善空間。

綜上所述，系統在常見摘要問題診斷、單一問題聚焦及逐步修訂引導方面已展現穩定表現。然而，對於複合問題、主觀文體判斷及語意邊界較模糊的問題類型，仍有提升空間。未來可進一步改善提示設計與問題類型判定機制，以提升診斷一致性及回饋品質。

系統回應情形（可複選）	發生次數
系統回應無明顯缺失	127
系統未辨識出本輪合理或主要的摘要問題	14
摘要已接近完成，但系統仍要求修改	8
同一輪要求學習者修改多個問題	7
回覆不自然、機械或難懂	2

表 6: 系統回應缺失次數統計

5 結論與未來展望

本研究針對現有摘要回饋系統較缺乏修訂順序規劃與形成性引導機制的問題，設計一套基於大型語言模型之書籍摘要形成性回饋系統。系統將摘要品質問題區分爲誤解、邏輯、遺漏、主觀、多餘及破碎鬆散六類，並依據由內容理解至語篇表達的修訂原則安排回饋順序，同時透過單一問題聚焦與鷹架式引導，協助學習者逐步完成摘要修訂。專家評估結果顯示，系統在問題辨識、問題定位、提供引導、教練語氣及最終答案揭露等面向皆獲得良好評價，並能於多數情況下維持單一問題聚焦與逐步修訂引導。整體結果初步支持本研究所建置之大型語言模型書籍摘要形成性回饋系統，作爲數位閱讀寫作教練之可行性。

然而，本研究仍有一些限制。首先，本研究以專家評估作爲初步驗證，尚未直接觀察學習者實際使用情形，因此無法推論其對摘要品質、修訂能力或閱讀理解之學習成效。其次，本研究僅邀請一位國文教師進行評估，評估結果仍可能受到個人教學觀點影響。此外，本研究正式評估僅採用 Gemma 4 31B 模型，未記錄跨模型比較。因此，目前結果僅能反映該模型與回饋架構組合下的完整系統表現，尚無法區分系統效能來自回饋架構或特定模型本身的能力。最後，系統在複合問題情境、完成狀態判定，以及主觀、多餘等語意邊界較模糊的問題類型上，仍可能出現診斷不一致或過度修正現象。

未來可在相同評估條件下，比較不同大型語言模型套用本研究回饋架構後的表現，釐清模型與各項回饋設計對系統表現的貢獻。另外，可將系統實際導入閱讀與寫作課程，蒐集學習者互動歷程與修訂資料，並透過實驗設計檢驗其對摘要品質、修訂策略及閱讀理解之影響。此外，未來也可增加書籍類型與評估樣本，並進一步改善複合問題辨識、完成狀態判定及回饋策略，以提升系統診斷一致性與大型語言模型於閱讀與寫作教學中的應用價值。

致謝

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

References

Veronika Barkela and Miriam Leuchter. 2024. Effectiveness of Automated Formative Feedback in an Online Tutorial for Promoting Summarizing. *Journal of Educational Technology Development and Exchange*, 17(1):67–95.

Robert-Mihai Botarleanu, Mihai Dascalu, Laura K. Allen, Scott Andrew Crossley, and Danielle S. McNamara. 2022. Multi-task Summary Scoring with Longformers. In Maria Mercedes Rodrigo, Noburu Matsuda,

Alexandra I. Cristea, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, volume 13355, pages 756–761. Springer International Publishing, Cham.

Christian Burkhart, Andreas Lachner, and Matthias Nückles. 2020. Assisting students' writing with computer-based concept map feedback: A validation study of the CohViz feedback system. *PLOS ONE*, 15(6):e0235209.

Chih-Hung Chen, Gwo-Jen Hwang, and Chin-Han Tsai. 2014. A progressive prompting approach to conducting context-aware learning activities for natural science courses. *Interacting with Computers*, 26(4):348–359.

Yutthasak Chuenchaichon. 2022. The Problems of Summary Writing Encountered by Thai EFL Students: A Case Study of the Fourth Year English Major Students at Naresuan University. *English Language Teaching*, 15(6):15.

Lester Faigley and Stephen Witte. 1981. Analyzing Revision. *College Composition & Communication*, 32(4):400–414.

Steve Graham, Michael Hebert, and Karen R. Harris. 2015. Formative Assessment and Writing: A Meta-Analysis. *The Elementary School Journal*, 115(4):523–547.

John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research*, 77(1):81–112.

Qi-Zhen Huang, Hou-Chiang Tseng, and Yao-Ting Sung. 2025. 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)*, pages 153–166, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Thorben Jansen, Jennifer Meyer, Johanna Fleckenstein, Andrea Horbach, Stefan Keller, and Jens Möller. 2024. Individualizing goal-setting interventions using automated writing evaluation to support secondary school students' text revisions. *Learning and Instruction*, 89:101847.

Walter Kintsch and Teun A. Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394.

Nitin Madnani, Jill Burstein, John Sabatini, and Tenaha O'Reilly. 2013. Automated Scoring of a Summary-Writing Task Designed to Measure Reading Comprehension. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–168, Atlanta, Georgia. Association for Computational Linguistics.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Wesley Morris, Scott Crossley, Langdon Holmes, Chaohua Ou, Mihai Dascalu, and Danielle McNamara. 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.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Yasuyo Sawaki, Yutaka Ishii, and Yoko Oi. 2024. The efficacy of automated feedback on summary content in an online formative assessment module for undergraduate L2 writing instruction in Japan. *Studies in Language Assessment*, 13(2):38–72.

Valerie J. Shute. 2007. FOCUS ON FORMATIVE FEEDBACK. *ETS Research Report Series*, 2007(1).

Tamara Sladoljev-Agejev and Jan Šnajder. 2017. Using Analytic Scoring Rubrics in the Automatic Assessment of College-Level Summary Writing Tasks in L2. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 181–186, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Nancy Sommers. 1980. Revision Strategies of Student Writers and Experienced Adult Writers. *College Composition & Communication*, 31(4):378–388.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained Summarization Evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2022. How to find strong summary coherence measures? a toolbox and a comparative study for summary coherence measure evaluation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6035–6049, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yao-Ting Sung, Chia-Ning Liao, Tao-Hsing Chang, Chia-Lin Chen, and Kuo-En Chang. 2016. The effect of online summary assessment and feedback system on the summary writing on 6th graders: The LSA-based technique. *Computers & Education*, 95:1–18.

John Sweller. 1988. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, 12(2):257–285.

David Wade-Stein and Eileen Kintsch. 2004. Summary Street: Interactive Computer Support for Writing. *Cognition and Instruction*, 22(3):333–362.

Paige Ware. 2011. Computer-Generated Feedback on Student Writing. *TESOL Quarterly*, 45(4):769–774.

Hiroyuki Yamanishi, Masumi Ono, and Yuko Hijikata. 2019. Developing a scoring rubric for L2 summary writing: a hybrid approach combining analytic and holistic assessment. *Language Testing in Asia*, 9(1):13.

李貞誼. 2021. 運用形成性評量回饋策略提升國中生閱讀能力及自我調整學習能力之準實驗研究. Master's thesis, 國立臺灣師範大學課程與教學研究所.

陳佳琳. 2013. 線上摘要評量與回饋系統之研究與應用. Master's thesis, 國立臺灣師範大學資訊教育學系.