From Uniform Feedback to Adaptive Support: AI-Driven Personalized Writing Feedback for Enhancing Self-Assessment Accuracy in Higher Education

ABSTRACT: 隨著生成式人工智慧技術的快速發展， LLM 逐漸應用於教育領域，特別是在自動化回饋與寫作評估方面。過去研究指出，形成性回饋能有效提升學生的自我評估能力與學習表現，然而，對於 LLM 所生成之回饋是否能根據學習者特質產生差異化效益，相關實證仍付之闕如。尤其在學生自我評估準確性(Self-Assessment Accuracy)上，LLM 與學習歷程間的交互作用尚未被系統性探討。本研究採準實驗設計，檢驗 LLM 生成之寫作回饋對學生 SAA 的影響，並進一步分析其對不同初始能力學生的調節效果。結果顯示，在整體樣本中，實驗組與控制組在後測 SAA 上無顯著差異；然而交互作用分析發現，回饋類型與初始 SAA 之間存在顯著效應。對於初始評估準確性較低的學生，LLM 回饋顯著提升其後測表現；而對 SAA 較高者則無顯著影響。這顯示 LLM 回饋對於低校準、學習處於劣勢的學生而言，可能是一種具成本效益且具擴展潛力的支持機制。本研究突顯 LLM 於提供補償性與適應性回饋上的潛能，特別是在學習者認知校準與自我監控能力方面具實際價值。同時亦呼籲未來 AI 回饋系統之設計，避免一體適用之策略，轉而納入動態調整與個別化機制，以實現真正具回應性的教學支持，推動人工智慧於教育評量中的深層轉化。

With the rapid advancement of generative artificial intelligence, large language models (LLMs) have become increasingly integrated into education, particularly for automated formative feedback and writing assessment. Prior research has shown that formative feedback can improve students' self-assessment accuracy (SAA) and learning outcomes. However, empirical evidence remains limited regarding whether AI-generated feedback can provide differentiated benefits based on learner characteristics. This study employed a quasi-experimental design to examine the effects of LLM-generated feedback on students' SAA and explore how these effects vary by initial ability. Results indicated no significant group-level difference in posttest SAA between the experimental and control groups. Nevertheless, interaction analysis revealed a significant effect between feedback type and initial SAA: students with lower baseline accuracy benefited more from AI-generated feedback, while students with higher baseline SAA showed no significant change. These findings suggest that LLM-generated feedback serves as a cost-effective and scalable support mechanism for underperforming learners, helping to reduce cognitive calibration gaps and promote metacognitive development. By embedding adaptive and personalized mechanisms within feedback systems, such AI-driven tools can advance equity and sustainability in higher education, especially in contexts where human feedback is limited. This study highlights the potential of intelligent education technologies to deliver responsive, differentiated support at scale, and advocates for a departure from one-size-fits-all approaches in AI-based assessment design to better align with the goals of sustainable and inclusive education.

Keywords: Self-assessment accuracy, Large language models, Formative feedback, AI-based feedback systems, AI-driven learning support

1.Introduction

隨著 AI 於教育領域中扮演日益關鍵的角色，AI 驅動的智慧學習系統逐漸被視為提升教育品質與公平的重要工具。這些系統不僅重塑學習方法，更為永續教育與大規模個別化支持提供創新解方 (Luckin, 2025)。為了在變動快速且自主性高度要求的環境中獲致良好學習成效，學生需展現高度的自我調節學習（Self-Regulated Learning,

SRL）。例如有效的時間與資源管理、個別設定學習目標，以及對策略與進度之持續性監控與評估(Chang et al., 2023)。而自我評估（Self-Assessment）成為 SRL 中不可或缺的一環。學生能否正確地判斷自己學習進度與表現，會直接影響後續學習策略的調整與資源分配。若學生對自身表現的判斷偏差過大，可能導致錯誤的學習決策，使學習成效大打折扣。

自我評估準確性（Self-Assessment Accuracy, SAA）被描述為學生對自身學習表現的主觀判斷與實際表現之間的一致程度(Wang et al., 2025)。作為 SRL 的主要構成，SAA 會直接影響學生是否能根據真實的學習狀況調整策略與目標。然而 Panadero et al., (2016) 指出：多數學生傾向高估自身表現，導致評估結果偏差，阻礙有效的學習調節。正因如此，越來越多研究主張應該採取具體且系統性的干預措施，以協助學生提升其自我評估的準確性(Luo & Zhou , 2024)。在眾多干預措施中，提供具針對性的 Feedback 被視為有效的干預策略，有助於促進高等教育中學生對任務要求與表現標準的理解，提高其自我評估的準確性 (Braumann et al., 2024)。表現較差的學生在 SAA 上通常落後於同儕，更需要依賴教師或 AI 系統提供的 Feedback。此時，具可擴展性與智慧化特徵的 AI 回饋系統（AI-driven feedback system），可協助教師針對學生的弱點提供即時、個別化的建議，不僅能提升 SAA，亦有助於學生更有效地調整學習策略與行為。對於提升學習公平性具有關鍵價值。

同時，這些學生透過具體與即時的 Feedback，可以獲得學習方向上的明確指引，有效修正錯誤認知並調整策略，從而促進其學習動力與元認知發展(Liu et al., 2025)。對於教師而言，提供 Feedback 是一項極具挑戰性的任務，尤其是針對複雜的寫作任務等涵蓋極高的時間與認知成本(Meyer et al., 2024)。隨著大型語言模型（Large Language Model, LLM）的迅速發展，AI 展現強化教學回饋的潛力，不僅能以更低的邊際成本提供即時、個人化的意見，也有助於教育資源弱勢環境中回饋機制的普及與永續性強化。例如 Meyer et al., (2024) 證明 LLM 生成的 Feedback 對學習成果具有正向影響，像是寫作表現、動機與情緒狀態的提升。

然而 LLM 所生成的 Feedback 是否能有效支持學生的 SAA 仍未獲得一致性證據。雖然 LLM 可以提供快速且一致的 Feedback，但 Feedback 在準確性、適切性與認知支持層面，與教師相比仍存在顯著差異 (Meyer et al., 2024)。LLM 所生成的 Feedback 可能缺乏情境敏感性，對於策略引導與錯誤識別的深度仍有待提升，進而影響其對 SAA 的潛在成效。Lew et al. (2010) 則提到，提供明確績效標準或具結構性的反饋對學生的 SAA 提升尤為關鍵，尤其是針對初始表現較差之學生，績效導向的 Feedback 能協助其重構自我理解與策略認知。

因此，本研究旨在探討 LLM 生成的 Feedback 是否能如同教師給予 Feedback 般有效促進學生的 SAA。透過實證驗證 AI 智慧系統是否能透過智慧化機制有效辨識學生初始 SAA 與學習表現差異，並據此提供具差異化與補償性的個別化建議，以了解其在推動具回應性與永續性的教育評量架構上的潛力。研究設計採隨機對照實驗，檢視學生在接收 LLM 回饋後 SAA 的變化趨勢與學習績效之關聯。進一步，為了探究 LLM 之 Feedback 對於不同學生群體的影響，研究亦納入學生初始表現與初始 SAA 表現作為調節變項，期望建構一種具差異化與智慧適應特徵的 AI 回饋支持模式，提升其於大規模教育環境中推廣的實用性與永續性。本研究具體問題如下：

- RQ1：LLM 所生成的 Feedback 是否能提高學生的 SAA？
- RQ2：與初始表現較高的學生相比，表現較差的學生是否可以通過 LLM 所生成的 Feedback 更好地提高他們的 SAA？
- RQ3：與初始 SAA 表現較高的學生相比，初始 SAA 表現較差的學生是否可以通過 LLM 所生成的 Feedback 更好地提高他們的 SAA？

2.Literature review

2.1 Empirical Foundations of Self-Assessment Accuracy and Feedback in Scalable Learning

Contexts

Self-assessment accuracy(SAA)又被稱為 calibration accuracy(Hacker & Bol, 2019)或 metacognitive monitoring accuracy(de Bruin & van Merriënboer, 2017)，描述學生自我評估表與自身實際表現結果相對應的程度。Self-assessment 涵蓋不同的技術與形式，協助學生自我監控學習歷程並判斷學習表現，進而促進學習調整與成效提升(Yan & Brown, 2017)。若學生能準確評估自身學習狀態，更能設定合理目標、精準監測進展，並針對學習策略做出有效調整(Rickey et al., 2025)。SAA 的研究受 SRL 所影響，被視為認知歷程中的關鍵組成。Self-assessment 活動本身亦已被證實能提升學生的反思能力與自我監控意識，是培養 SRL 技能的重要歷程之一(Andrade, 2019)。Thiede et al.,(2010)在文本理解領域的研究指出，SAA 較高的學生更能有效辨識需要重新學習的內容，因此能表現出更佳的學習成果。凸顯 SAA 不僅是一項元認知指標，更是實現有效學習的必要條件。具備高 SAA 學生，能更準確掌握自身學習狀況，針對弱點調整策略，有效提升學習成效。Ernst et al.,(2025)則指出高準確度的 SAA 能促進學生對自身學習過程的現實理解，有助於減少過度自信或錯誤判斷所導致的策略失誤，提升學習效能與動機。

然而，學生在 Self-assessment 出現不準確的情形已被研究多次提到(León et al., 2023; Panadero et al., 2016)。也突顯出有效介入措施學生 SAA 的重要性。依據檢索利用理論(Cue utilization theory)，學生會依賴與自身表現相關的線索進行自我評估，但由於線索品質不一，容易影響評估的準確性(Kakaria et al., 2024)。Koriat (1997)近一步細分出可預測實際表現的具診斷性線索(Diagnostic cues)與無法預測實際表現的非診斷性線索(nondiagnostic cues)。例如學生會根據自身閱讀速度(歸類於 Nondiagnostic cues) 來判斷對自身對文章的理解程度，從而做出錯誤的評估：而當學生收到教師給予的文本修訂 Feedback 時(歸類於 Diagnostic cues)，這些高品質的 Feedback 才能幫助學生進行較為真實的自我評估。

學生需要被提供有效的 Diagnostic cues，才能提升 SAA(Winstone et al., 2017)。這也凸顯 Feedback 在學生 SAA 上傳遞 Diagnostic cues 的重要功能(Butler & Winne, 1995; Panadero et al., 2016)。Gutierrez de Blume(2022)在統合分析中指出，Feedback 有助於提升學生自我評估的準確性，具中等程度效果。於此同時，關於學生水準與 Feedback 之間交互作用的研究仍相對稀少(Maier & Klotz, 2025)。過往研究在探討 SSA 對於 Feedback 時，往往忽略 SAA 本身亦可能成為 Feedback 效應的調節因子。若學生初始 SAA 較低，代表學生在校準與監控上存在明顯困難，此時所接收的結構化回饋可作為校準參照點，有助於辨識學生學習落差並進行策略性修正(Ernst et al., 2025; Nederhand, et al., 2019)。這亦呼應了 diagnostic feedback 在支持高風險學習者自我監控與調節中的核心作用(Wille et al., 2025)。探討學生的初始表現與 SAA 如何共同調節 Feedback，不僅有助於釐清 Feedback 機制運作之精細路徑，亦對高等教育中個別化教學介入具實質啟發意義。

2.2 Feedback Literacy and Self-Assessment as Synergistic Components in Intelligent Systems

反饋素養(Feedback literacy) 是決定學生是否能夠進行有意義的 self-assessment 並從中受益的重要因素。雖然本研究不足以全面重新詮釋 Feedback literacy，但我們仍嘗試將 Feedback literacy 視為一項與 self-assessment 密切相關的能力來進行討論。Feedback literacy 的範疇除了涵蓋對評價資訊的詮釋，亦包含回應 Feedback 時的情緒管理與將 Feedback 內化為學習資源的互動性理解(Molloy et al., 2020)。Carless and Boud (2018)將 Feedback literacy 界定為學習者所需具備的理解、技能與傾向，使其能夠有效理解回饋資訊，並據此強化自身的學習或工作策略。後續研究則近一步探討 Feedback literacy 融入課程設計的三個核心面向：主動尋求相關資訊、有效處理回饋內

容，以及根據回饋採取調整行動(Malecka et al., 2020)。Nicol (2021)則提出內部反饋(Internal Feedback)的概念，主張學生在 Feedback 歷程中應發展自我生成的認知評估機制，以深化 Feedback 的理解與應用。

儘管 Feedback literacy 與 self-assessment 本質上是不同的概念(Kang et al., 2025)，但兩者在自我調節和終身學習等實現教育的核心目標上都發揮著重要作用(Boud, 1999; Winstone & Carless, 2019)。為了更好地理解兩者並有效應用於實務中，研究應採取整合性的觀點，並審視兩者之間的互動關係。在 self-assessment 的過程中，存在著許多發展 feedback literacy 的機會。具備良好 Feedback literacy 的學生，也更有可能進行更有意義的 self-assessment。與其他素養相同，Feedback literacy 是一個持續且漸進的過程。透過適當的設計，Feedback literacy 不僅能讓學生主動尋求外部回饋(External feedback)，還能透過將自身表現與各種參照資訊進行比較來產生 Internal Feedback 這種重複性的過程有助於提升學生的 feedback literacy。

然而，在 self-assessment 如何影響 Feedback literacy 的議題上，過去的研究並沒有過多琢磨。但 Feedback literacy 在 self-assessment 中的促進角色，可以從以下兩個層面進行談討。第一，self-assessment 不單單僅有「自我(self)」，在 self-assessment 中，「他人(others)」也存在同樣關鍵(Yan & Brown, 2017)。為促進以學習為中心的 Feedback，學生須主動尋找並整合具適切性的人際或環境資源的能力。Boud(1999)也提到，self-assessment 歷程需要學生主動從教師、同儕、家長等相關環境與人員尋求 feedback。具備良好 feedback literacy 的學生，更傾向主動尋求 feedback，也更了解專業性、可信度與人際關係等因素會如何影響回饋歷程與品質。更可能找到對 self-assessment 真正有幫助的回饋資訊(Malecka et al., 2022)。第二，self-assessment 中會產生 Internal Feedback。學生會將自己的作品與某些參照標準進行比較，從而產生 Internal Feedback。這些 Internal Feedback 能支援自我評估歷程中的多個面向，例如確立自我評估的標準、辨識優勢與弱點，以及調整當前的學習策略。具備 feedback literacy 的學生較能產生具備學習導向的高品質內在回饋，因此更有潛力有效運用自我評估的結果來促進自身的學習與改進(Yan, 2020)。

尋求回饋(feedback seeking)意指學生主動獲取與自身作品或學習表現相關的資訊。為被視為 feedback literacy 的核心構成，因其能將學習者的內部認知歷程與外部資訊資源有效連結。feedback seeking 可以分為兩種策略：主動詢問(Inquiry) 與觀察監測(Monitoring)。Inquiry 意指學生直接向他人請求對自己進展或已察覺問題的意見；Monitoring 則是指學生從環境中獲取資訊，例如比較自己過去的表現與他人表現、參考範例、評量標準，或查閱其他相關資源(Ashford & Cummings, 1983; Joughin et al., 2021; Leenknecht et al., 2019)。僅有 External feedback 並無法為學生帶來學習成效，唯有學生能透過處理與應用這些資訊，進一步產生 Internal Feedback，才能真正促進學習成長。也因此 Internal Feedback 是 Feedback literacy 的另一個核心行為要素。學生自行發展洞見從 Feedback 中建構意義，並將其應用於未來的行動調整。不論回饋的來源為何，學生都必須透過自身的認知系統加以詮釋與篩選後才能真正運用。

2.3 The Transition from Human Feedback to AI-Based Intelligent Feedback Systems

形成性評量(formative assessment) 旨在持續調整教學內容，以符合學生的需求(Filsecker & Kerres, 2012) 儘管有部分學者提出爭議(Bennett, 2010)。formative assessment 仍被認為是促進有效學習最具影響力的因素之一。其中 formative feedback 便是 formative assessment 的核心要素之一。formative feedback 是一種提供給學習者的資訊，其目的是協助他們調整思維歷程與行為策略，進而促進學習表現的提升(Shute, 2008)。根據 Hattie and Timperley(2007)的研究，有效的 Feedback 需要涵蓋三個問題：Feed Up、Feed Back 與 Feed Forward。Feed Up 涉

及學生明確陳述學習目標、建立清楚的方向與目標感。Feed Back 聚焦於學生當下的表現，並在脈絡中指出錯誤，幫助學生理解自己的學習狀況並辨識具體可改善的區域。Feed Forward 則提供未來行動的指引。而這三個問題又可以應用在四個層次上：task level、process level、self-regulation level 與 self level(Hattie & Timperley, 2007)。

除了內容層面的因素，文字上語氣、語言清晰度、prompt 的適切使用與文字長度也對 feedback 的有效性具有關鍵影響。feedback 的語氣應具有鼓勵性，同時避免過度正向(Kluger & DeNisi, 1996)。這樣的做法能建立支持性的學習環境，並確保回饋具有建設性(Brookhart, 2017)。平衡的語氣能激勵學生，同時保留對其成長至關重要的批判性意見。過於籠統的 feedback 文字會形成學生理解上的阻礙，因此 Feedback 應具體、直接且易於理解(Ossenberg et al., 2019)。在文字長度方面，Kulhavy et al., (1985)提出 Feedback 應力求簡潔明確。但 Van der Kleij et al., (2015) 在其統合分析中提出，相較於過度簡化的回饋，內容詳盡的回饋可提升學習成效。

隨著計算機發展，基於自動化、電腦的回饋也逐漸在教育研究領域當中被證實具有實質效益。其中一大特性便是能產生即時回饋，對於學習表現較差的學生而言，即時回饋已被證實具有良好成效(Shute, 2007)。然而，一對一輔導因為能針對學生做出個別化調整，被視為最有效的學習形式，而計算機系統仍難以比擬此種高適應性學習支持。相較於計算機系統，基於 LLM 的系統能提供更高程度的互動性，從而實現更具適應性與參與感的學習體驗。同時也有較少偏見的認知優勢。此外，自動化回饋還有較少偏見的認知優勢(Zawacki-Richter, et al., 2019)。Nguyen et al., (2023)在數學教育研究中評估 LLM 處理學生對小數概念錯誤的回饋效能，發現其在提供恰當回饋方面整體表現良好。Seßler et al., (2023)在寫作教學方面使用 LLM 生成即時回饋，並佐證 LLM 建設性評論支持學生寫作發展上的實用性。Gabbay and Cohen(2024)在程式設計教育領域中提到，LLM 雖能有效偵測 MOOC 課程中的程式作業錯誤，但在提供精確或可行的回饋建議方面仍有不足。Estévez-Ayres et al., (2024)則提到，LLM 在評估涉及並行錯誤(concurrency errors)的練習題時能力有限，凸顯其理解複雜程式概念的挑戰。Koutcheme et al., (2024)則觀察到，LLM 在初學者程式課程中傾向給予過度正面的回饋，可能忽略一些應被關注的重要問題。

在較早的文獻綜合描述中，Zhai et al., (2020)檢視了基於 AI 的科學評量研究，並聚焦於技術實作、效度與教學特性三個層面，發現大多數研究偏重於效度，較少處理技術與教學面向的議題。雖然近兩年已有一些研究嘗試探索 LLM 在科學教育評量中的應用。例如 Wu et al., (2023)應用預訓練 BERT 模型搭配 zero-shot prompting 對學生書寫作答進行評分，展現 LLM 在評分任務中的可行性。Latif & Zhai(2024)比較微調過的 BERT 模型與 GPT-3.5 在自動評分學生答案的表現，發現 GPT-3.5 在評分準確度上顯著優於 BERT。Guo et al., (2024)發展一套用以針對科學教育題項生成回饋的 multi-agent system。但該研究存在幾項重要限制：第一，研究並未將系統所產生的回饋與真實教師的回饋進行比較，因此與實際課堂教學的契合度尚未驗證。第二，研究評估面向僅關注過度讚美(over-praise)與過度推論(over-inference)，忽略了明確性(clarity)與具體性(specificity)等其他關鍵教學品質指標。第三，缺乏針對多元課堂情境的實證驗證，導致其在教育場域中的普遍適用性仍受質疑。LLM 系統的確為學習回饋帶來新的可能，但其若欲取代或補足人類導師在教學中的角色，仍需面對語用層次、教育倫理與實作成效等多重挑戰。唯有在技術效能與教育價值之間取得平衡，才能真正實現科技輔助學習的深度轉化。

3.系統架構

3.1 Dataset Analysis

本研究所使用之資料集來自校內通識教育中心國文課程學生於 2024 年 1 月至 2025 年 7 月間所撰寫之文章，共計 7158 篇。這些文章涵蓋議論文和敘述性論文，能有效反映一般大學生在自然語境下的中文寫作表現。本研究依據文章主題與任務性質，將資料集劃分為 8 個群集，各群集代表不同的文本類型，分別強調語義組織、論證邏輯或敘事技巧等面向，呈現多元且具代表性的語篇結構。此分類方式有助於語言模型在訓練回饋生成時，學習辨識並對應不同文本的內容特性與寫作需求。*Figure 1* 與 *Figure 2* 分別代表訓練集文章的文章長度與文章字數，模型需具備足夠的語境處理能力，以理解篇章上下文並生成具針對性的語言回饋。

Figure 1: Distribution of essay lengths

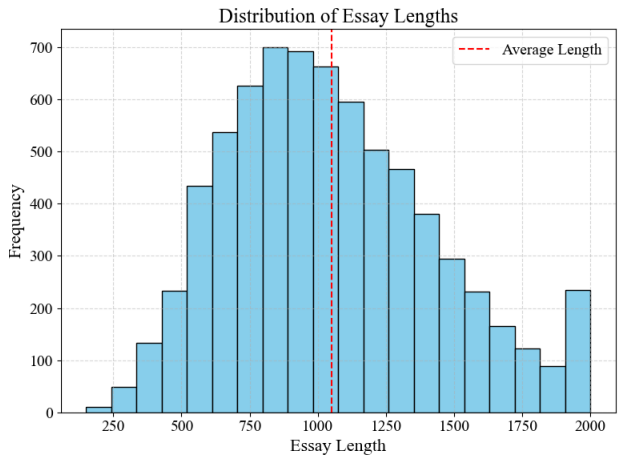
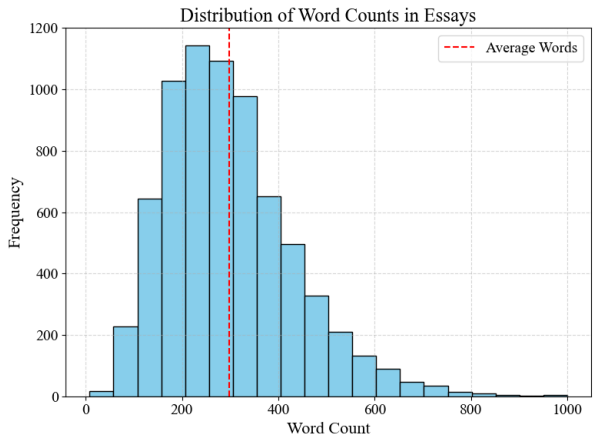


Figure 2: Distribution of essay words

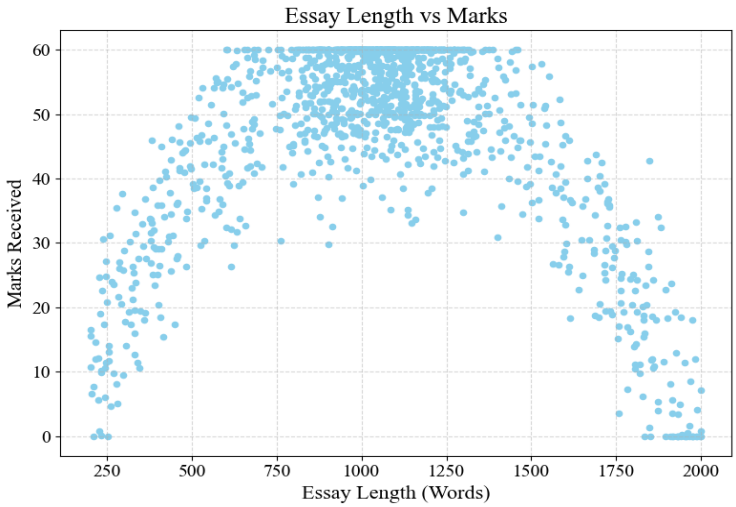


為了檢視語言特徵對得分的影響，Equation(1)計算了各篇章的詞彙豐富度(Type-Token Ratio, TTR)。TTR 是衡量一段文本中詞彙多樣性的基本指標(Richards, 1987)，反映文本中不同詞的數量與總詞數的比例。其中 Word types 表示不重複的詞數，Word tokens 表示所有詞出現的總數。

$$\text{TTR} = \frac{\text{Number of word types}}{\text{Number of word tokens}} \tag{1}$$

TTR 的分佈結果如 *Figure 3* 所表示，整體樣本的 mean 為 0.34，SD 為 0.06。顯示大多數學生在詞彙使用上具有中等程度的多樣性。TTR 與得分之間雖非完全線性相關，但得分較高之論文往往伴隨較高詞彙變化度，顯示語彙多樣性可能與文本說服力與內容發展程度相關。

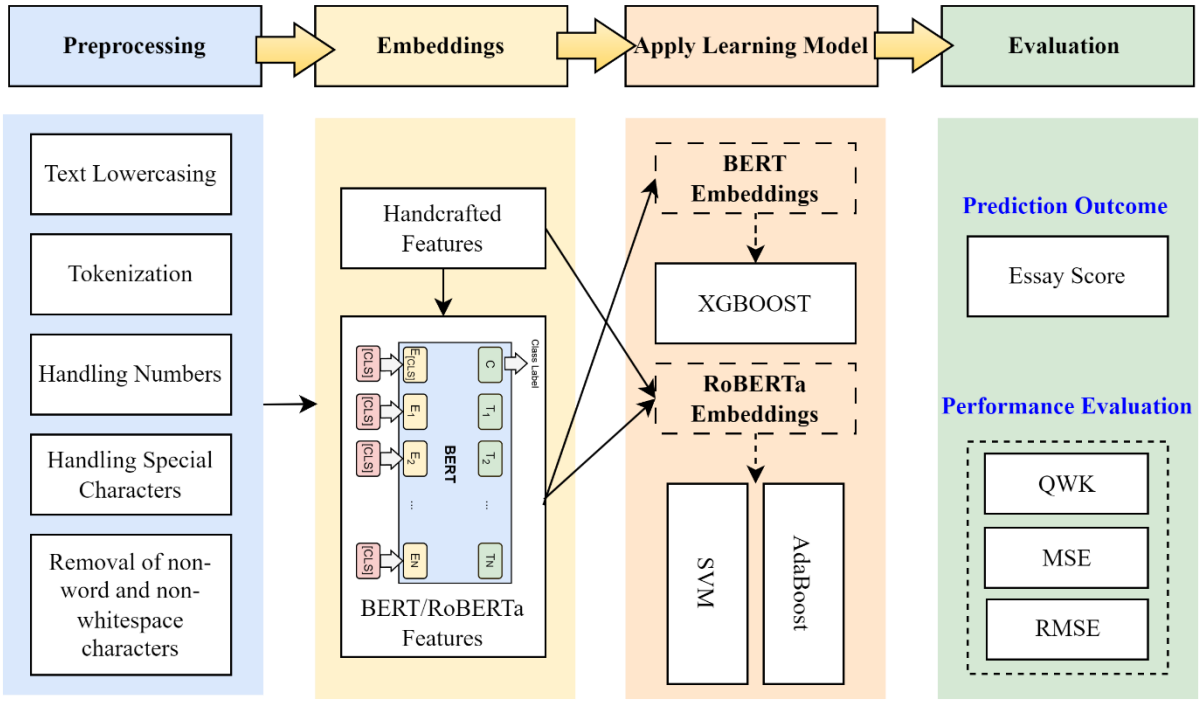
Figure 3: Average sentence length vs. marks



3.2 Overview of the Model

本研究所開發之寫作評估系統整體架構如 Figure 4 所表示。在預處理(Preprocessing)中會執行一些操作，例如 Text Lowercasing、Tokenization、Handling Numbers、Handling Special Characters、Removal of non-word and non-whitespace characters 等，確保文本數據被清理、標準化，並準備好進行特徵提取和模型訓練。如果 Preprocessing 未被正確執行，則可能導致模型偏誤，降低提取特徵的品質，進而降低模型的預測準確率。不當的斷詞、特殊符號處理失當或缺少標點符號，都會扭曲文本涵義，並削弱後續 Embeddings 的效能。為了處理上述問題，本研究失實多種錯誤處理機制，例如在 Preprocessing 進行驗證，確保僅有有效詞彙得以進入後續階段；對於缺失或錯誤資料，會以占位符替代或應用填補技術(imputation methods) 處理；數字或特殊符號這類邊緣案例會被移除或正規化，以便後續模型使用。

Figure 4: Layered architecture of proposed model.



在語意相似度上，我們使用預訓練的 BERT 模型來為學生文章與訓練文本生成 embeddings，並計算向量之間的餘弦相似度(cosine similarity)以獲取語意相似度分數，分數介於 0 至 1 之間。對於文本結構，系統藉由搜尋「however」、「therefore」等過渡詞，以及句間過渡語來判斷文章的連貫性。Equation(2)為詞彙豐富度指標，有助於理解作者的語言能力與認知成熟度。

$$\text{Vocabulary Richness} = \frac{\text{Number of Unique Words}}{\text{Total Number of Words}} \tag{2}$$

所有特徵計算完成後，會以欄方式(column-wise)串接進原始資料集中，並針對文章重複此過程，直到全部完成整合。接下來 RoBERTa 模型會提取語境化的 embeddings，捕捉文章中精細的語意資訊。RoBERTa 具備有效編碼語言結構與語意關聯的能力，透過 Handcrafted Features 與 RoBERTa 結合，得以掌握文章的結構與語意品質，提升寫作評估系統的準確性。

3.3 BERT

BERT(Bidirectional Encoder Representations from Transformers)能捕捉文本中複雜的語境關係與細微的語意差異，顯著提升文章評分的準確性。其雙向架構(bidirectional architecture)可同時理解上下文，幫助模型推敲句意、辨識語言線索、維持文本的一致性，進而更準確地表達文章內容。

由於 BERT 是在大規模語料庫上預訓練，具備理解各種語言慣例與寫作風格的能力。BERT 所產生的語境化嵌入向量(contextualized embeddings)能整體呈現作文的內容，增強模型對作文品質的準確判斷與深入分析能力。

在資料輸入模型前，我們會先提取手工特徵(Handcrafted Features)，接著與 BERT 串接為最終的特徵向量。其公式如 Equation(3) (4) (5)：

$$x_i^{\text{hand}} = \phi_{\text{hand}}(e_i) \tag{3}$$

$$x_i^{\text{BERT}} = \phi_{\text{BERT}}(e_i) \tag{4}$$

$$x_i = \left[x_i^{\text{hand}} \mid x_i^{\text{BERT}} \right] \tag{5}$$

其中x_i^{hand}為第 i 篇文章之的手工特徵向量，x_i^{BERT}為 BERT 模型對第 i 篇作文產生的嵌入向量，x_i為合併後的特徵向量，ϕ 則為特徵提取函數。而模型訓練階段的公式如 Equation(6)，模型使用參數Θ對每篇文章的合併特徵向量x_i預測其得分$\hat{y}_i$。

$$\hat{y}_i = g(x_i; \Theta) \tag{6}$$

3.4 RoBERTa

RoBERTa(Robustly Optimized BERT Approach)是 BERT 的改良版本，旨在克服 BERT 模型的一些限制。也因為 RoBERTa 是在更大規模的資料集與更長的輸入序列上所訓練，在自動文章評分(Automated Essay Scoring, AES)這類需要理解局部與全局語境的任務中表現更加出色。Equation(7) (8) (9)為 RoBERTa 的特徵向量整合公式。在本研究中，RoBERTa 被作為獨立模型使用，其接收 RoBERTa embeddings 與 Handcrafted Features 作為輸入，例如語法錯誤數量、語意相似度與、詞彙豐富度。

$$x_i^{\text{hand}} = \phi_{\text{hand}}(e_i) \tag{7}$$

$$x_i^{\text{RoBERTa}} = \phi_{\text{RoBERTa}}(e_i) \tag{8}$$

$$x_i = \left[x_i^{\text{hand}} \mid x_i^{\text{RoBERTa}} \right] \tag{9}$$

3.5 Evaluation Metrics

Evaluation Metrics 在寫作評估系統中扮演衡量模型效能的角色。Evaluation Metrics 能以量化方式評估模型在精確性、一致性與可靠性等層面的表現，從而判斷模型對文章內容與品質的評估能力是否符合預期。寫作評估系統使用 QWK(Quadratic Weighted Kappa)作為主要效能指標，可以用來衡量模型預測分數與人工標註分數之間的一致性。公式如 Equation(10)所表示。其中O_{ij}為觀察到的分數矩陣(實際與預測的混淆矩陣)， E_{ij}為期望分數矩陣(假設完全隨機)，ω_{ij}為權重矩陣，κ則是所有可能的分數等級數。相較於僅計算分類正確率的傳統準確度指標，QWK 能夠納入評分等級的序數性特徵，並對預測值與實際值接近程度加權考量，提供更具區辨力的模型評估依據。為了更全面檢視模型在連續變數預測任務中的表現，研究分別納入 MSE 與 RMSE 輔助指標。公式如 Equation(11)與 Equation(12)所表示。其中y_i為真實文章分數，$\hat{y}_i$為預測分數，n是文章總數。

$$\kappa = 1 - \frac{\sum_{i,j} \omega_{ij} O_{ij}}{\sum_{i,j} \omega_{ij} E_{ij}} \tag{10}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{12}$$

4. Method

4.1 Study sample

本研究的參與者為來自台灣南部某大學工程學院的 64 位大學生(N=64)，資料收集時間為 2024 年 4 月至 7 月。該實驗課程屬於校內人文素養課程的一部分。為確保學生在既有學業背景與基礎能力上的同質性，研究採用分層隨機分派法，將學生分為實驗組(EG，N = 32)與控制組(CG，N = 32)。在整體樣本中，男生佔 56.75%，平均年齡為 18.2 歲；女生占 44.25%，平均年齡為 18.7 歲。

4.2 Study design and procedure

實驗流程圖如 *Figure 5* 所表示。整體實驗透過電腦以線上問卷的形式進行。實驗開始前會有一個簡短說明。說明結束後，兩組學生皆開始進行一篇繁體中文的論說文(argumentative essay)。時間上限為 50 分鐘。論說文題目為以下兩者之一，無論是哪一種題目，學生都需要針對題目表達個人意見，並加以說明與論證。題目一：「你是否同意以下說法？一個人的成功來自選擇而非天賦。請舉出具體理由與例子來支持你的觀點」。題目二：「是否同意以下說法？AI 終將取代大多數人類的工作。請舉出具體理由與例子來支持你的觀點」。

在完成第一版初稿後，所有學生須進行一次自我評估前側，以評估自己的寫作表現。隨後，實驗組需要藉由本研究開發之 LLM 系統生成 Feedback，並根據 Feedback 進行修訂。LLM 系統所提供的指示如下：「請根據系統所提供之 Feedback(如 *Table 1* 所表示) 修訂您的文章，盡可能地進行完善修正，請花足夠的時間進行修訂。」而控制組雖然同樣有機會修訂自己的文章，但 LLM 系統並不會給予任何具體的回饋內容。相反的，控制組所看到的是一段標準化提示語：「請再次閱讀您的文章，並盡可能進行修訂。請花足夠的時間進行修正。」這段提示的目的僅是為了讓控制組學生在沒有實際回饋的情況下，也能參與修訂過程，以保持兩組在修訂程序上的一致性。兩組的修訂的過程中，皆可以隨時查看原始任務說明與自身的初稿，並直接進行修改。學生修訂時間上限同樣為 50 分鐘。兩組之間的唯一差異在於：實驗組有實際的回饋表與修訂指引，而控制組僅有一般性指令。

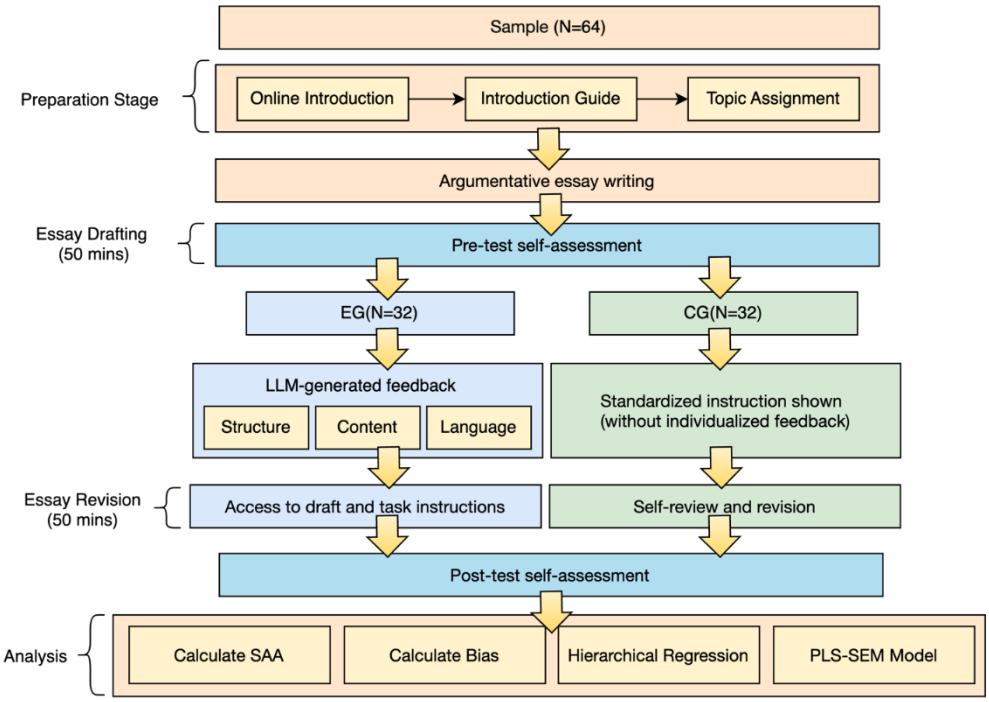
Table 1: Feedback example generated by LLM 系統

面向	改進提示	具體改進建議
structure	文章缺乏清楚的開頭與結尾，使得整體架構不明，讀者較難掌握內容邏輯。建議在	補上引言與結論段落。內容安排上應將相關觀點整理於同一段落中，並使用轉折或因果類語句(如：

	段落之間加入銜接語句，以提升段落之間的連貫性。	「然而」、「因此」、「另一方面」)協助引導閱讀。
content	文章中的論點支持力道不足，缺乏明確且具體的例子來佐證主張，且立場表達不夠清晰，內容顯得籠統。	增加具體事例來強化論點說服力，並清楚表達文章立場。避免使用模糊或概括性語句，讓讀者更容易理解作者的觀點。
language	文章中存在多處拼字或文法錯誤，詞彙使用較為單一。部分句子過長，建議進行斷句與重構，以提升可讀性。	仔細檢查並修正拼寫與語法錯誤，嘗試使用更豐富的字彙。將冗長句子拆解成較簡潔的語句，並加入如「然而」、「此外」、「儘管如此」等過渡語來強化語意連貫。

在完成文章修訂後，兩組皆需要再次進行自我評估後測，評估自己修訂後的寫作表現。實驗結束時，研究團隊亦蒐集了學生的基本背景資料，包括年齡、性別與大學國文總成績。整個實驗耗費三節課(共 150 分鐘)完成。

Figure 5: Experimental procedure



4.3 Measures

學生在完成自我評估前側與後測候，會根據 Likert 五點量表(0 = 非常差，6 = 非常好) 對自己的寫作表現進行自我評估，題目為：「請評估你剛才完成的這篇文章，你認為這篇文章的品質如何」。研究以自我評估分數與 LLM 系統給予分數之間的絕對差距來衡量 SAA。其公式參考自 Schraw(2009)公式(Equation(13))，其中 N 為學生樣本數，c_i為第 i 個學生的自我評估分數；p_i 為第 i 個學生的 LLM 系統分數。

$$\text{Absolute Accuracy Index} = \frac{1}{N}\sum_{i=1}^{N}(c_i - p_i)^2 \tag{13}$$

同時，本研究也評估學生的 bias，即自我評估與實際表現之間的簡單差值(正值表示高估，負值表示低估)。依據 Schraw(2009)公式(Equation(14))，納入 bias 是為了了解學生自我評估準確性的變化情形。透過分析學生的 bias 分數，我們希望探討學生 SAA 的進步是否可解釋為對自身表現之高估或低估的減少。其中 N 為學生樣本數，$c_i - p_i$為學生的自評分數減去 LLM 系統分數。然而，bias 分數並不等於絕對準確性(Absolute Accuracy Index)，故在解讀結果時可能具誤導性(Schraw, 2009)。

$$\text{Bias Index} = \frac{1}{N}\sum_{i=1}^{N}(c_i - p_i) \tag{14}$$

學生在完成自我評估前側與後測候，會根據 Likert 五點量表(0 = 非常差，5 = 非常好) 對自己的寫作表現進行自我評估，題目為：「請評估你剛才完成的這篇文章，你認為這篇文章的品質如何」。研究以自我評估分數與 LLM 系統給予分數之間的絕對差距來衡量 SAA。其公式參考自 Schraw(2009)公式(Equation(15)，其中 N 為學生樣本數，c_i為第 i 個學生的自我評估分數；p_i 為第 i 個學生的 LLM 系統分數。

$$\text{Absolute Accuracy Index} = \frac{1}{N}\sum_{i=1}^{N}(c_i - p_i)^2 \tag{15}$$

同時，本研究也評估學生的 bias，即自我評估與實際表現之間的簡單差值(正值表示高估，負值表示低估)。依據 Schraw(2009)公式(Equation(16)，納入 bias 是為了了解學生自我評估準確性的變化情形。透過分析學生的 bias 分數，我們希望探討學生 SAA 的進步是否可解釋為對自身表現之高估或低估的減少。其中 N 為學生樣本數，$c_i - p_i$為學生的自評分數減去 LLM 系統分數。然而，bias 分數並不等於絕對準確性(Absolute Accuracy Index)，故在解讀結果時可能具誤導性(Schraw, 2009)。

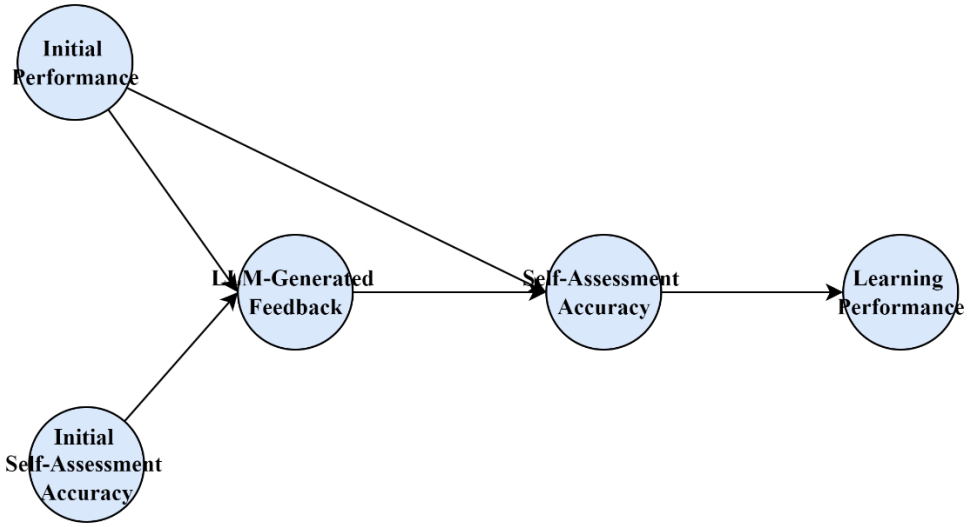
$$\text{Bias Index} = \frac{1}{N}\sum_{i=1}^{N}(c_i - p_i) \tag{16}$$

4.4 Analytic approach

本研究使用 Smart-PLS 4 的 structural equation model(SEM)進行分析。PLS-SEM 適用於處理具有中介與調節變項之複雜模型，並能同時處理形成性與反映性潛在構念，亦適合進行預測導向研究。SEM Model 如所 *Figure 6* 表示，為了檢驗 LLM 生成 Feedback 對 SAA 的影響(RQ1)，SEM 設計 LLM-Generated Feedback(LLMF)至 Self-Assessment Accuracy(SAA)的直接路徑。並使用 Bootstrapping 重抽樣法檢驗路徑係數之統計顯著性。

對於了解初始表現(Initial Performance, IP)與初始自我評估準確度(Initial Self-Assessment Accuracy, ISAA)對 LLM 所產生的回饋(LLMF)，本研究進一步納入兩個交互作用項：(1) IP 至 LLMF 的直接路徑(RQ2)；(2)ISAA 至 LLMF 的直接路徑(RQ3)。為降低多重共線性問題，所有連續變項皆先行標準化後再建立交互作用項。同時，本研究亦以 SPSS 進行階層迴歸分析作為補充驗證，依序建立三層模型以觀察模型解釋力與效果變化。Model1 僅納入組別變項以檢驗主效應。Model2 加入 IP 與 ISAA 兩個前測控制變項。Model3 再納入兩個交互作用項以檢驗調節效應。

Figure 6: Proposed Structural Equation Model



5. Results

5.1 Descriptive statistics

Table 2 與 *Figure 4* 呈現 SEM 中各面向之 means, standard deviation, 與 partial correlation。學生的 IP 平均得分為 2.43(*SD* = 0.85)，ISAA 平均得分為 1.82(SD = 0.73)，兩者存在一定落差。在使用寫作評估系統後，學生的 LLMF 得分為 3.01(*SD* = 0.64)，SAA 也相較於 ISAA 有所提升(*M* = 2.10, *SD* = 0.81)，而最終 Learning Performance 平均為 2.89(SD = 0.76)。

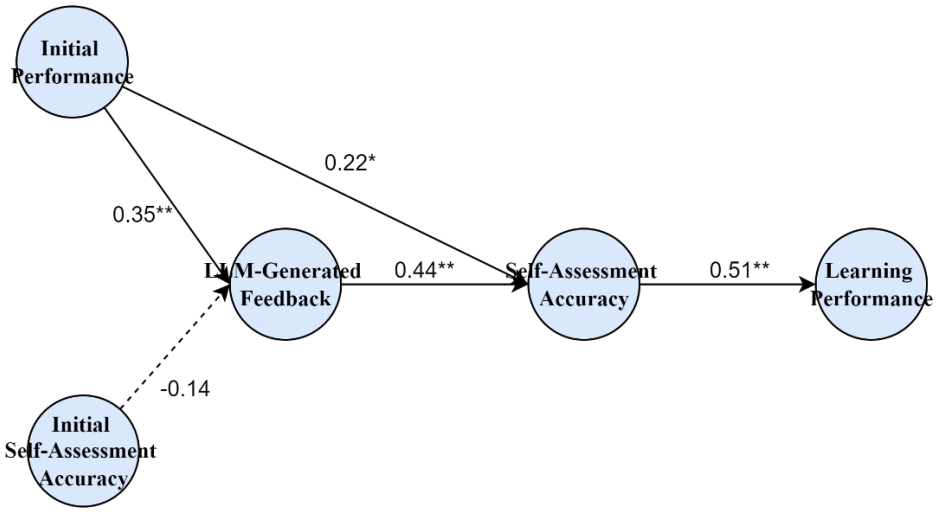
在 partial correlation 上，IP 與 LLMF(*r* = 0.35, *p* <.01**)及 IP 與 LP(*r* = 0.40, *p* <.01**)呈現正向偏相關，表示起始能力較佳的學生，傾向獲得更多回饋且學習表現也相對較高。ISAA 與 SAA 呈現正向偏相關(*r* = 0.39, *p* <.01**)，顯示初始評量準確度較高者，傾向在學習後保有較佳的自我評估能力。LLMF 與 SAA(*r* = 0.44, *p* <.01**)以及 LLMF 與 LP (*r* = 0.36, *p* <.01**)皆呈現中度正向偏相關，指出 LLM 所產生的回饋可能對自我評估的修正與學習成果具有實質助益。SAA 與 LP 間的偏相關係數最高(*r* = 0.51, *p* <.01**)，顯示更準確的自我評估能力與更佳的學習表現間存在穩定的正向關係。

Table 2: Means, standard deviations, and partial correlations

Variable	*Mean*	*SD*	*IP*	*ISAA*	*LLMF*	*SAA*	*LP*
Initial Performance (IP)	2.43	0.85					
Initial Self-Assessment Accuracy (ISAA)	1.82	0.73	-0.28**				
LLM-Generated Feedback (LLMF)	3.01	0.64	0.35**	-0.14			
Self-Assessment Accuracy (SAA)	2.10	0.81	0.22*	0.39**	0.44**		
Learning Performance (LP)	2.89	0.76	0.40**	-0.10	0.36**	0.51**	

Note. *p < .05; **p < .01; ***p < .001

Figure 4: Path coefficients of the structural model



Note. Solid lines represent significant paths (*$p < .05$; **$p < .01$; ***$p < .001$); dashed line represents a nonsignificant path.

5.2 Testing hypotheses

為了檢驗本研究假設，我們進行三層次的階層迴歸分析，結果如表 *Table 3* 與 *Table 4* 所表示。在 Model1 僅納入組別變項，用以檢驗 LLMF 對 SAA 的直接影響(RQ1)。結果顯示，LLMF 的主效應並未達顯著($\beta = 0.10, p = .098$)，並未支持 RQ1。在 Model2 中，基於 Model1 並加上 IP 與 ISAA 兩個前測變項。結果顯示，IP($\beta = 0.17, p = .011$)與 ISAA($\beta = 0.43, p < .001$***)皆為 SAA 顯著正向預測因子，模型解釋力亦顯著提升($\Delta R^2 = 0.12$)。在 Model3 進一步納入兩個交互作用項：LLMF * IP 以及 LLMF * ISAA，檢驗回饋是否根據學生的前測表現產生不同調節效果。LLMF * IP 的交互作用則未達顯著($\beta = 0.07, p = .243$) 不支持假設 RQ2a；LLMF * ISAA 的交互作用顯著($\beta = 0.22, p < .01$**)支持假設 RQ2b。而 Model3 總解釋力為 $R^2 = .35$，顯著優於 Model2($\Delta R^2 = 0.12$)。為更清楚呈現交互作用結果，*Figure 7* 與 *Figure 8* 顯示了群組與前測自評準確度(SAA)及 bias 之間的交互趨勢圖。*Figure 7* 顯示在 SAA 較高的情況下，實驗組學生的後測準確度提升幅度大於控制組，呈現正向交互作用。*Figure 8* 則說明偏差程度與後測偏差呈現的趨勢差異。

Table 3: Hierarchical Regression Analysis Predicting Students' Self-Assessment Accuracy

Predictor	*B*	*SE*	*β*	*t*	*p*
LLMF (EG = 1)	0.10	0.06	0.10	1.67	.098
Initial Performance(IP)	0.18	0.07	0.17	2.57	<.05*
Initial Self-Assessment Accuracy (ISAA)	0.41	0.05	0.43	8.20	<.001***
LLMF * IP	0.07	0.06	0.07	1.17	.243
LLMF * ISAA	0.22	0.06	0.22	3.67	<.01**

Note. *$p < .05$; **$p < .01$; ***$p < .001$

Table 4: Hierarchical Regression Model Summary Predicting Students' Self-Assessment Accuracy

Model	*R²*	*ΔR²*	*F(df)*	*ΔF*	*p*

Model 1	0.17		10.28(1, 63)		
Model 2	0.29	0.12	18.94(1, 63)	15.3	<.001***
Model 3	0.35	0.06	19.76(1, 63)	8.20	<.001***

Note. Model 1 僅納入組別變項。Model 2 加入 Initial Performance 與 Initial Self-Assessment Accuracy。Model 3 進一步納入兩組交互作用項(LLMF * IP 與 LLMF * ISAA)

*p < .05; **p < .01; ***p < .001

Figure 7: Interaction between Pre SAA and Group

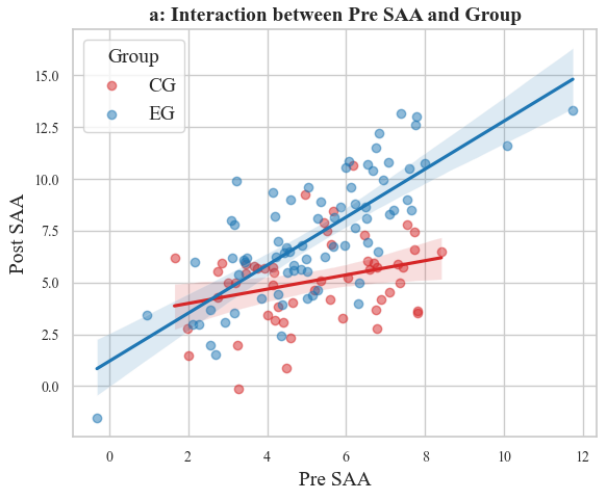
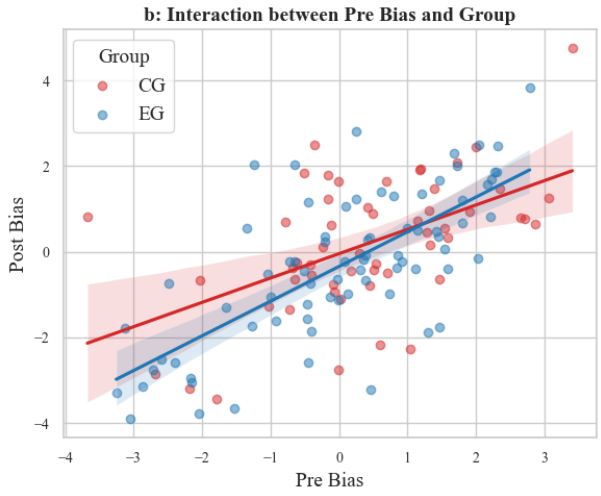


Figure 8: Interaction between Pre Bias and Group



Note. SAA 數值愈低代表評估愈準確；Bias 為負代表低估，為正則代表高估。

6. Discussion

針對 LLM 所生成的 Feedback 對於 SAA 的影響上，實驗組雖然在後測 SAA 平均上升，但與控制組之間的差異並未達顯著水準。此結果凸顯出，LLM 回饋並非對所有學生均產生等效的正向影響。與 Hattie and Timperley (2007) 所指出的「回饋效果高度仰賴學生解讀與處理回饋的能力」的觀點一致。也呼應 Winstone et al., (2017) 對於回饋接受歷程所需條件的提醒。即便 LLM 能提供形式上符合準則的 feedback，若學生無足夠的 feedback literacy，也難以有效轉化為內化的自我監控行為(Carless & Boud, 2018)。

我們進一步探討 Initial Performance (IP)是否會調節 LLM 回饋對 SAA 的影響。研究結果指出，Initial Performance (IP)與 LLM-Generated Feedback (LLMF)之間的交互並未達顯著。Initial Performance (IP)的高低並未顯著影響學生從 LLM 回饋中受益的程度。學生即便表現中等偏低，若具備較強的反思與校準能力，仍能進行有效的自我修正。相對地，高表現者若缺乏反思機制，亦可能對回饋反應遲鈍(Lew et al., 2010)。

然而，Initial Self-Assessment Accuracy (ISAA)與 LLM-Generated Feedback (LLMF)之間呈現顯著交互作用。初始 SAA 較低的學生，在接受本研究 LLM 系統回饋後，後測 SAA 明顯提升。此結果與 Butler & Winne (1955)、Koriat (1997)之結果呼應。學生若能從外部獲得具「校準參考性」的指引，將有助於辨識認知偏差並進行策略修正。特別是 SAA 偏低的學生，通常缺乏有效監控標準與策略評估能力，因此更仰賴高品質外部回饋以進行自我調整(Ernst et al., 2025)。

整體而言，本研究針對 LLM 生成的 Feedback 在學生自我評估歷程中提供了細緻的實證證據。LLM 生成回饋並非對所有學生皆產生一致性的正向效果，而是在考量個體差異後，其效果才被具體呈現。這種回饋形式如同一種「針對性的支持工具」，顯著提升那些原先傾向高估自己表現學生的 SAA。同時，本研究拓展了當前 LLM Feedback 的研究脈絡，特別是在探討自我評估效果時納入個體差異的觀點(Meyer et al., 2024; Panadero et al., 2016)。即便 Feedback 是由 LLM 自動提供，它仍能即時地為自我評估能力較弱的學生提供個別化的支持。在實際應用 LLM Feedback 機制時，應避免採取「一體適用」的設計邏輯，應該依照學習者特性，審慎建構具差異化與適性化的回饋模式，以真正發揮 LLM 在教育中的促進功能。

7. Conclusion

本研究透過實證實驗，檢驗由 LLM 所生成之 Feedback 對學生 SAA 的影響，並進一步探討其對不同學習者族群的調節效果。研究結果顯示，雖然 LLM 所生成的回饋在整體樣本中並未展現顯著效果，但對於初始 SAA 較低的學生而言，LLM 回饋能有效促進其評估準確性，具有補償性與適應性的潛力。此結果不僅驗證了個別差異在回饋介入中的關鍵角色，更突顯 LLM 在智慧教學系統中的潛在價值。特別是在支援高風險學習者調整學習策略與強化自我監控歷程方面的應用前景。透過內嵌差異化診斷與回饋機制的 AI 回饋系統，能在大規模教育場域中，以低成本實現個別化支持，有助於推動教育公平性與長期永續發展。

然而，本研究在實作上仍遇到些許限制。第一個主要限制可能來自於我們所設計的回饋方式，儘管本研究是參考過往實證研究與理論基礎所制定。但本研究所提供的 Feedback 是擴充性回饋（elaborated feedback），並未包含任何可歸類的結果知識，可能對那些原本自我評估就較為準確的學生產生負面影響。未來研究應進一步探討影響 feedback 對學生 SAA 有效性的具體特徵。第二個限制是介入時間較短，僅進行為期 150 分鐘的實驗。雖然短時間的研究能減少外在混淆變項的干擾，但也限制了學生在課後進行延伸性自我評估與反思的機會。未來研究仍需探討這種 SAA 的提升是否具有長期持續性，並更深入了解學生在更長的時間與多次回饋循環中，如何處理與內化所獲得的回饋。第三個限制是控制組設計相對較弱，控制組的學生在文本修訂過程中並未收到任何回饋，研究難以全面評估由 LLM 生成回饋的相對效益。未來研究可以將 LLM 回饋與來自其他來源（例如教師或其他自動化系統）的類似回饋進行比較，以評估我們的研究結果在多大程度上是由 LLM 特有的特性所驅動，而這些特性本身具有某些優勢與限制。此外，後續研究也應進一步探討此類系統於不同學科、教育階段與文化脈絡中的跨場域適用性，提升其作為智慧教育基礎建設的實務價值。

在理論貢獻上，本研究回應並延伸了過往關於 formative feedback、feedback literacy 與自我評估互動關係的討論，強調高品質回饋在校準認知偏差、強化元認知歷程上的作用。相較於以往研究多聚焦於教師或靜態回饋形式，本研究提供了 LLM 作為動態回饋工具的初步實證，對 AI 輔助學習評估領域具有開創意義。同時，本研究也為智慧教育科技如何以學習者中心的方式提供差異化支持，建立了實證基礎，有助於未來更有效地設計 AI 教育介入策略。

而在實務貢獻上，研究結果指出 LLM 生成 feedback 不應採取通用模式，而應依據學習者的初始能力與評估準確性進行差異化配置。未來設計此類 AI 回饋系統時，應納入個別診斷機制與動態調整策略，以提供具針對性的教學支持。具備智能推理與決策模組的回饋系統，未來將可依據學生歷程自動調整回饋層級與語氣，實現真正意義上的智慧化、回應式教育。當 LLM 系統能結合學習歷程資料與個人化引導，將更有可能實現真正具回應性的學習回饋環境，並作為高等教育中智慧型學習平台的一環，推動規模化、永續且公平的教育創新。

References

Andrade, H. L. (2019, August). A critical review of research on student self-assessment. In *Frontiers in education* (Vol. 4, p. 87). Frontiers Media SA. https://doi.org/10.3389/feduc.2019.00087

Ashford, S. J., & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational behavior and human performance*, *32*(3), 370-398. https://doi.org/10.1016/0030-5073(83)90156-3

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, *8*(2-3), 70-91. https://doi.org/10.1080/15366367.2010.508686

Boud, D. (1999). Avoiding the traps: Seeking good practice in the use of self assessment and reflection in professional courses. *Social work education*, *18*(2), 121-132. https://doi.org/10.1080/02615479911220131

Braumann, S., van de Pol, J., Kok, E., Pijeira-Díaz, H. J., van Wermeskerken, M., de Bruin, A. B., & van Gog, T. (2024). The role of feedback on students' diagramming: Effects on monitoring accuracy and text comprehension. *Contemporary Educational Psychology*, *76*, 102251. https://doi.org/10.1016/j.cedpsych.2023.102251

Brookhart, S. M. (2017). *How to give effective feedback to your students*. Ascd.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, *65*(3), 245-281. https://doi.org/10.3102/00346543065003245

Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Chang, D. H., Lin, M. P. C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability*, *15*(17), 12921. https://doi.org/10.3390/su151712921

de Bruin, A. B., & van Merriënboer, J. J. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction*, *51*, 1-9. https://doi.org/10.1016/j.learninstruc.2017.06.001

Ernst, H. M., Prinz-Weiß, A., Wittwer, J., & Voss, T. (2025). Discrepancy between performance and feedback affects mathematics student teachers' self-efficacy but not their self-assessment accuracy. *Frontiers in Psychology*, *15*, 1391093. https://doi.org/10.3389/fpsyg.2024.1391093

Estévez-Ayres, I., Callejo, P., Hombrados-Herrera, M. Á., Alario-Hoyos, C., & Delgado Kloos, C. (2024). Evaluation of LLM tools for feedback generation in a course on concurrent programming. *International journal of artificial intelligence in education*, 1-17. https://doi.org/10.1007/s40593-024-00406-0

Filsecker, M., & Kerres, M. (2012). Repositioning Formative Assessment from an Educational Assessment Perspective: A Response to Dunn & Mulvenon (2009). *Practical Assessment, Research & Evaluation*, *17*(16), n16.

Gabbay, H., & Cohen, A. (2024, July). Combining LLM-generated and test-based feedback in a MOOC for programming. In *Proceedings of the eleventh ACM conference on learning@ scale* (pp. 177-187). https://doi.org/10.1145/3657604.3662040

Guo, S., Latif, E., Zhou, Y., Huang, X., & Zhai, X. (2024). Using generative AI and multi-agents to provide automatic feedback. *arXiv preprint arXiv:2411.07407*. https://doi.org/10.48550/arXiv.2411.07407

Gutierrez de Blume, A. P. (2022). Calibrating calibration: A meta-analysis of learning strategy instruction interventions to improve metacognitive monitoring accuracy. *Journal of Educational Psychology*, *114*(4), 681.

Hacker, D. J., & Bol, L. (2019). Calibration and self-regulated learning: Making the connections. https://doi.org/10.1017/9781108235631.026

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112. https://doi.org/10.3102/003465430298487

Joughin, G., Boud, D., Dawson, P., & Tai, J. (2021). What can higher education learn from feedback seeking behaviour in organisations? Implications for feedback literacy. *Assessment & Evaluation in Higher Education*, *46*(1), 80-91. https://doi.org/10.1080/02602938.2020.1733491

Kakaria, S., Simonetti, A., & Bigne, E. (2024). Interaction between extrinsic and intrinsic online review cues: perspectives from cue utilization theory. *Electronic Commerce Research*, *24*(4), 2469-2497. https://doi.org/10.1007/s10660-022-09665-2

Kang, C., Huang, J., Liu, Y., & Yin, H. (2025). Development and validation of a generic self-assessment scale for K-12 teachers as feedback givers: Insights from item response theory and factor analysis. *Humanities and Social Sciences Communications*, *12*(1), 1-10. https://doi.org/10.1057/s41599-025-04927-4

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, *119*(2), 254.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of experimental psychology: General*, *126*(4), 349. https://doi.org/10.1037/0096-3445.126.4.349

Koutcheme, C., Dainese, N., Sarsa, S., Hellas, A., Leinonen, J., & Denny, P. (2024). Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4-as-a-judge. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1* (pp. 52-58). https://doi.org/10.1145/3649217.3653612

Kulhavy, R. W., Lee, J. B., & Caterino, L. C. (1985). Conjoint retention of maps and related discourse. *Contemporary Educational Psychology*, *10*(1), 28-37. https://doi.org/10.1016/0361-476X(85)90003-7

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Leenknecht, M., Hompus, P., & van der Schaaf, M. (2019). Feedback seeking behaviour in higher education: the association with

students' goal orientation and deep learning approach. *Assessment & Evaluation in Higher Education*, *44*(7), 1069-1078. https://doi.org/10.1080/02602938.2019.1571161

León, S. P., Panadero, E., & García-Martínez, I. (2023). How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy. *Educational Psychology Review*, *35*(4), 106. https://doi.org/10.1007/s10648-023-09819-0

Lew, M. D., Alwis, W. A. M., & Schmidt, H. G. (2010). Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education*, *35*(2), 135-156. https://doi.org/10.1080/02602930802687737

Liu, C. C., Hwang, G. J., Yu, P., Tu, Y. F., & Wang, Y. (2025). Effects of an automated corrective feedback-based peer assessment approach on students' learning achievement, motivation, and self-regulated learning conceptions in foreign language pronunciation. *Educational technology research and development*, 1-22. https://doi.org/10.1007/s11423-025-10484-z

Luckin, R. (2025). Nurturing human intelligence in the age of AI: rethinking education for the future. *Development and Learning in Organizations: An International Journal*, *39*(1), 1-4. https://doi.org/10.1108/DLO-04-2024-0108

Luo, R. Z., & Zhou, Y. L. (2024). The effectiveness of self-regulated learning strategies in higher education blended learning: A five years systematic review. *Journal of Computer Assisted Learning*, *40*(6), 3005-3029. https://doi.org/10.1111/jcal.13052

Maier, U., & Klotz, C. (2025). Students ignore their mistakes: Elaborated error feedback processing in a digital learning system. *Contemporary Educational Psychology*, 102395. https://doi.org/10.1016/j.cedpsych.2025.102395

Malecka, B., Boud, D., & Carless, D. (2022). Eliciting, processing and enacting feedback: mechanisms for embedding student feedback literacy within the curriculum. *Teaching in Higher Education*, *27*(7), 908-922. https://doi.org/10.1080/13562517.2020.1754784

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (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. https://doi.org/10.1016/j.caeai.2023.100199

Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education*, *45*(4), 527-540. https://doi.org/10.1080/02602938.2019.1667955

Nederhand, M. L., Tabbers, H. K., & Rikers, R. M. (2019). Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. *Applied Cognitive Psychology*, *33*(6), 1068-1079. https://doi.org/10.1002/acp.3548

Nguyen, H. A., Stec, H., Hou, X., Di, S., & McLaren, B. M. (2023, August). Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European conference on technology enhanced learning* (pp. 278-293). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42682-7_19

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in higher education*, *46*(5), 756-778. https://doi.org/10.1080/02602938.2020.1823314

Ossenberg, C., Henderson, A., & Mitchell, M. (2019). What attributes guide best practice for effective feedback? A scoping

review. *Advances in Health Sciences Education*, *24*(2), 383-401. https://doi.org/10.1007/s10459-018-9854-x

Panadero, E., Brown, G. T., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational psychology review*, *28*, 803-830. https://doi.org/10.1007/s10648-015-9350-2

Richards, B. (1987). Type/token ratios: What do they really tell us?. *Journal of child language*, *14*(2), 201-209. https://doi.org/10.1017/S0305000900012885

Rickey, N., Panadero, E., & DeLuca, C. (2025). How do students self-assess? examining the metacognitive processes of student self-assessment. *Metacognition and Learning*, *20*(1), 1-29. https://doi.org/10.1007/s11409-025-09430-4

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and learning*, *4*, 33-45. https://doi.org/10.1007/s11409-008-9031-3

Seßler, K., Xiang, T., Bogenrieder, L., & Kasneci, E. (2023, August). Peer: Empowering writing with large language models. In *European conference on technology enhanced learning* (pp. 755-761). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42682-7_73

Shute, V. J. (2007). Focus on formative feedback. *ETS Research Report Series*, *2007*(1), i-47. https://doi.org/10.1002/j.2333-8504.2007.tb02053.x

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189. https://doi.org/10.3102/0034654307313795

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*, *47*(4), 331-362. https://doi.org/10.1080/01638530902959927

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research*, *85*(4), 475-511. https://doi.org/10.3102/0034654314564881

Wang, W. S., Lin, C. J., Lee, H. Y., Huang, Y. M., & Wu, T. T. (2025). Enhancing self-regulated learning and higher-order thinking skills in virtual reality: the impact of ChatGPT-integrated feedback aids. *Education and Information Technologies*, 1-27. https://doi.org/10.1007/s10639-025-13557-x

Wille, E., Opheim, H. M. S., Kisa, S., & Hjerpaasen, K. J. (2025). Building Resilience and Competence in Bachelor Nursing Students: A Narrative Review of Clinical Education Strategies. https://doi.org/10.20944/preprints202506.2171.v1

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational psychologist*, *52*(1), 17-37. https://doi.org/10.1080/00461520.2016.1207538

Winstone, N., & Carless, D. (2019). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge. https://doi.org/10.4324/9781351115940

Wu, X., He, X., Liu, T., Liu, N., & Zhai, X. (2023, June). Matching exemplar as next sentence prediction (mensp): Zero-shot prompt learning for automatic scoring in science education. In *International conference on artificial intelligence in education* (pp. 401-413). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_33

Yan, Z. (2020). Self-assessment in the process of self-regulated learning and its relationship with academic achievement. *Assessment & Evaluation in Higher Education*, *45*(2), 224-238. https://doi.org/10.1080/02602938.2019.1629390

Yan, Z., & Brown, G. T. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education*, *42*(8), 1247-1262. https://doi.org/10.1080/02602938.2016.1260091

Yan, Z., & Brown, G. T. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education*, *42*(8), 1247-1262. https://doi.org/10.1080/02602938.2016.1260091

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *International journal of educational technology in higher education*, *16*(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, *56*(1), 111-151. https://doi.org/10.1080/03057267.2020.1735757