Developing an AI-empowered Chinese Composition Assessment System for Hong Kong Students

Hiu Laam Naomi Lee[1], Yik Elinor Wong[2], Wai Yin Koey Chung[1], Chi Fuk Henry So[*3]
[1,3] Centre for Learning, Teaching and Technology, [2] Department of Chinese Language Studies, [3] Department of Mathematics and Information Technology
The Education University of Hong Kong
[*] hcfso@eduhk.hk

Abstract: *With the advancements in natural language processing and the growing adoption of Large Language Models (LLMs) in education, AI-based automated essay evaluation has emerged as a significant topic in research on writing literacy assessment. Despite this, there is limited research on Chinese automated essay scoring systems, leaving the subject largely underexplored. This paper introduces an AI-empowered Chinese composition assessment system specifically designed for primary and secondary schools in Hong Kong. The system features customizable prompts based on marking criteria, employs multiple AI models to enhance grading accuracy, and enables rapid report generation. We will outline the system's design, workflow, and future directions for field testing and eventual implementation.*

Keywords: Automated Essay Evaluation, AI in education, Chinese composition assessment, educational technology

1. Introduction

Writing literacy is an essential skill that enhances one's ability to express ideas and communicate effectively. It is vital for students' lifelong development and capacity to adapt to societal changes in the future (Genlott & Grönlund, 2013). With advancements in natural language processing (NLP) technology, AI-based automated essay evaluation has become a prominent area of research in assessing writing literacy. The repetitive task of evaluating students' writing requires a significant investment of time and effort from teachers. Consequently, automated essay evaluation significantly alleviates the grading workload for educators. This project aims to develop an AI-empowered Chinese composition assessment system for Hong Kong primary and secondary school teachers and students. The system evaluates writing proficiency through automated analysis of compositions written by native Chinese students and provides targeted feedback for improvement. The system features several key elements:

1.Customized Experience: Each participating school will have a unique experience with no waiting time. Teachers can adjust the system's requirements and customize mark allocations to fit their teaching schedules and marking schemes. The system accommodates any topics and prerequisites, automatically generating prompts based on the specified criteria for immediate use by the AI models.

2.Enhanced Grading Accuracy: The system employs multiple AI models to assess assignments, which improves grading accuracy. Two AI models will collaborate to evaluate each assignment, while a third model will act as a separate verifier if the first two models produce significantly different scores.

3.User-Friendly Interface and Rapid Feedback: A dedicated interface for teachers will be established as the system develops. Teachers can easily upload compositions and receive comprehensive AI-generated reports within minutes, all within a unified application framework.

This paper reviews the need for an AI-empowered composition assessment system in Hong Kong. It provides an overview of the design and development of the system, highlighting the integration of pre-built AI models, particularly

focusing on the prompt framework and workflow associated with the system. Lastly, the paper concludes by discussing the upcoming field testing and the anticipated limitations of the system.

2. The use of AI and NLP in educational assessments

The study of automated essay assessment began in 1966 with Page's pioneering research on the Project Essay Grader system (Page, 1966). This innovative system utilized a multiple regression program to replicate the evaluations of human markers. It predicted essay scores by analyzing factors such as the frequency of uncommon words, the use of prepositions and commas, and the overall length of the essay. The program then compared these scores with those of randomly selected essays, and surprisingly, it was found to be indistinguishable from the assessments made by English teachers. Since then, automated essay scoring (AES) has remained an important area of research (Ke & Ng, 2019).

The advent of generative artificial intelligence models that can comprehend and produce human language marks a significant advancement in the application of AI for educational assessments. Prior to the emergence of pre-built AI models such as ChatGPT and Gemini, AES systems were primarily developed using traditional machine learning techniques. Since the 1990s, these systems have focused on identifying patterns within various features derived from extensive datasets (Ramesh & Sanampudi, 2022).

Past achievements in AES include notable systems such as the Intelligent Essay Assessor (IEA) by Foltz et al. (1999), e-rater V2 by Attali and Burstein (2006), and IntelliMetric by Rudner et al. (2006). These systems are characterized as handcrafted feature AES, whose effectiveness primarily depends on the quality of the features designed by experts. In contrast, more recent systems developed over the last decade employ neural network models that automatically learn features for essay scoring, as highlighted by Hussein et al. (2019). A significant advancement in this field was introduced by Alikaniotis et al. (2016), who presented a deep neural network (DNN) model capable of autonomously extracting essay scoring features without relying on predefined templates. Recently, DNN-based models have garnered considerable attention. Uto (2021) has compiled a list of new models from 2016 to 2021 that underscores this emerging trend.

The introduction of ChatGPT in 2022, followed by the releases of Gemini and Claude in 2023, has dramatically enhanced the accessibility of AI language models. According to Mizumoto and Eguchi (2023), these AI language models hold significant potential as tools for AES. Furthermore, Latif and Zhai (2024) highlight the importance of conducting additional research on the capabilities of various generative AI models in the context of automatic scoring in education.

3. The current needs of the system in Hong Kong

Over the past few decades, numerous AES platforms have been developed, primarily aimed at assessing the writing skills of second language (L2) learners. Most of these platforms are tailored for evaluating English writing, with significantly fewer systems available for other languages, especially Chinese. Consequently, the potential of LLMs in the domain of Chinese AES remains largely untapped. In Hong Kong, only a limited number of studies have examined automated essay evaluation using datasets derived from local students. Notably, these datasets are exclusively composed of English writings (Chan et al., 2023; Lee et al., 2009). We resonate with other researchers in the field and acknowledge the potential of leveraging prebuilt LLMs for automated essay evaluation. Our concept for the AI-empowered Chinese Composition Assessment System is groundbreaking in its application of prebuilt LLM platforms for automated grading. Moving forward, we will explore the feasibility of customizing assessment criteria to generate prompts on the user side, allowing the system to provide essay scores and reports within a minute.

4. Development of the AI-empowered Chinese Composition Assessment System

The AI-empowered Chinese Composition Assessment System is designed to assess writing proficiency through automated analysis of student compositions, delivering personalized developmental advice using pre-trained AI models.

This system is supported by a prebuilt LLM studio from the Centre for Learning, Teaching, and Technology (LTTC) at The Education University of Hong Kong (EdUHK). It offers a Model-as-a-Service (MaaS) platform that enables developers to tailor intelligent workflows for educational applications.

The system is designed to incorporate an intuitive application interface that facilitates user-friendly navigation. Users can configure assessment criteria as needed and upload their compositions on a page-by-page basis. The development of the system and application interface is currently underway. Figure 1 shows a detailed illustration of the design and structure of the AI-empowered Chinese Composition Assessment System, highlighting its innovative features and capabilities.

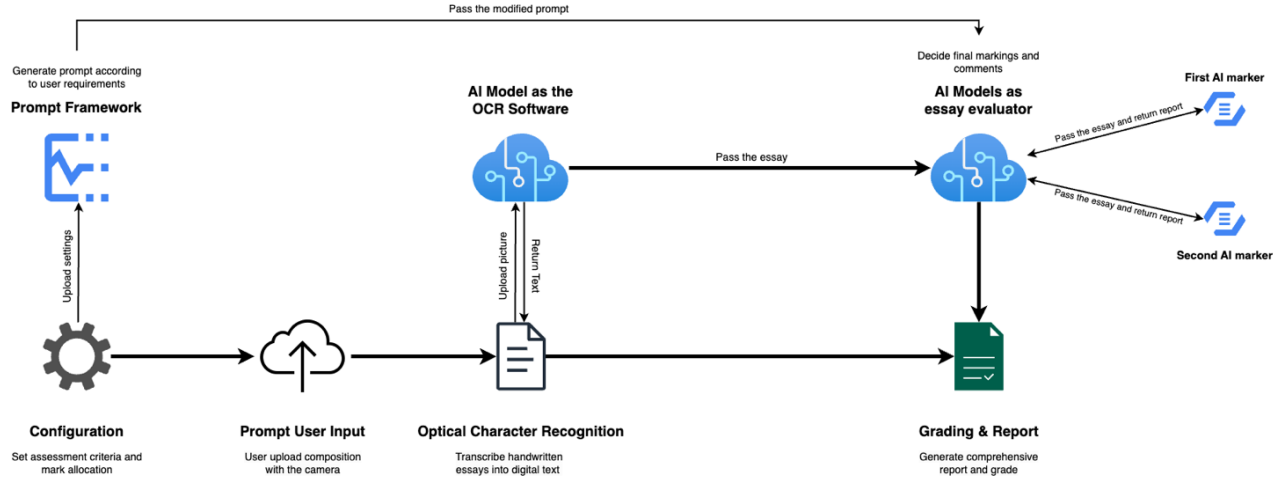

Fig.1 The structure of the AI-empowered Chinese Composition Assessment System

類別	框架	只可以自訂黃色格內容		
#任務	#角色	你是一名香港	中學　　中文	科老師。
		我會提供有關課堂的資料，作業的細節以及評分準則。		
	#指令	你需要根據所提供的指令認真思考並作出評分。繁體字和簡體字相通，並不當作有錯別字。如果一個詞中部分是簡體字，部分是繁體字，則簡體部分算錯別字。		
#規則	#技能	###技能1: 評價作文	根據上傳的學生作文，總結並理解作文內容並做出整體評價。	
		###技能2: 提高作文能力	根據寫作能力學習重點提出建議。舉例說明如何寫作。	
		###技能3: 反饋及評價	提供正面反饋並讚賞學生優點。	
		###技能4: 批注及修改	根據評分的結果，對學生作文中細節做批注。	
		###技能5: 符號，字體，字數	評估標點符號和錯別字並給出建議。	
	#回覆格式	1: 使用段落格式評價學生論文。		
		2: 使用評分標準以表格形式做出評估，在每項評分後根據評分重點說明評分的依據。		
		3: 引用學生作文部分論文原文，並對此部分原文做出批注和改善建議。		
		4: 使用繁體中文進行回覆。		
#內容	科目	為	中文	
	作文提目	為	投訴無牌小販擺賣	
	體裁	為	應用文	
	類型	為	投訴信	
	年級	為	中二	
	字數限制	為	沒有	
	總分	為	34	
#總分	內容	的總分為	16	
	格式	的總分為	15	
	標點字體	的總分為	3	
	錯別字	每個扣分	0.5	最多扣　3　分
	...	的總分為	...	
	...	的總分為	...	
#評分指引	###內容:	內容部分總分16分：表明寫信原因，得2分；交代最少2項投訴原因，每項投訴原因，根據描述的清晰程度、合理程度評分，每項滿分為4分，共計8分；提出要求收件人交代或解釋或處理投訴狀況，得4分；提供自己聯絡方法供處理部門聯絡自己，得2分。		
	###格式:	格式部分總分15分：啟首語寫上「敬啟者」，得3分；結束語在正文後的新一行靠左空兩格寫上「此致」，得3分；在結束語「此致」的下一行靠左頂格寫上題目要求的受文者，得3分；在受文者下一行靠右寫上「投訴人」及投訴人的姓名，加上啟告語「啟」或「謹啟」，得3分；最後一行靠左頂格寫上日期，包括年、月、日，得3分。		
	###標點字體:	按能否適切使用標點符號、適當地分段、字體秀麗整齊程度綜合評分，最高3分。如有寫字，最少給1分。		
	...	...		

Fig.2 Prompt Framework

The prompt generation framework for the system will be grounded in expert consultations and assessment rubrics established by the Hong Kong Education Bureau. Teachers can input their specific requirements, including but not limited to the subject, composition topic, text type, genre, grade level, word limit, and total marks. Furthermore,

educators can customize the marking criteria based on various aspects, such as content, format, punctuation, spelling errors, or any elements they wish to prioritize. Figure 2 depicts the concept of prompt generation within the backend of the system, with the customizable elements highlighted in yellow. The requirements entered by users will be processed to generate prompts that will be supplied to the AI model. Figure 3 showcases examples of these prompts.

指令（第一部分）	#角色 你是一名香港中學中文科老師。我會提供有關課堂的資料，作業的細節以及評分準則。你需要根據所提供的指令認真思考並作出評分。 繁體字和簡體字相通，並不當作有錯別字。如果一個詞中部分是簡體字，部分是繁體字，則簡體部分算錯別字。 #技能 ###技能1: 評價作文 根據上傳的學生作文，總結並理解作文內容並做出整體評價。 ###技能2: 提高作文能力 根據寫作能力學習重點提出建議。舉例說明如何寫作。 ###技能3: 反饋及評價 提供正面反饋並讚賞學生優點。 ###技能4: 批注及修改 根據評分的結果，對學生作文中細節做批注。 ###技能5: 符號，字體，字數 評估標點符號和錯別字並給出建議。 #回覆格式 1: 使用段落格式評價學生論文。 2: 使用評分標準以表格形式做出評估，在每項評分後根據評分重點說明評分的依據。 3: 引用學生作文部分論文原文，並對此部分原文做出批注和改善建議。 4: 使用繁體中文進行回覆。
指令（第二部分）	#內容 科目為中文。 作文提目為投訴無牌小販擺賣。 體裁為應用文。 類型為投訴信。 年級為中二。 字數限制為沒有。 總分為34。 #總分 內容的總分為16。 格式的總分為15。 標點字體的總分為3。 錯別字每個扣分0.5。 最多扣3分。 #評分指引 ###內容: 內容部分總分16分：表明寫信原因，得2分；交代最少2項投訴原因，每項投訴原因，根據描述的清晰程度、合理程度評分，每項滿分為4分，共計8分；提出要求收件人交代或解釋或處理投訴狀況，得4分；提供自己聯絡方法供處理部門聯絡自己，得2分。 ###格式: 格式部分總分15分：啟首語寫上「敬啟者」，得3分；結束語在正文後的新一行靠左空兩格寫上「此致」，得3分；在結束語「此致」的下一行靠左頂格寫上題目要求的受文者，得3分；在受文者下一行靠右寫上「投訴人」及投訴人的姓名，加上啟告語「啟」或「謹啟」，得3分；最後一行靠左頂格寫上日期，包括年、月、日，得3分。 ###標點字體: 按能否適切使用標點符號、適當地分段、字體秀麗整齊程度綜合評分，最高3分。如有寫字，最少給1分。

Fig.3 Prompts generated for the AI model

The AI-empowered system is designed to transcribe handwritten essays into digital text and deliver a comprehensive report that includes grades and comments within one minute. This process involves various tasks, such as Optical Character Recognition (OCR) of handwritten work (refer to Figure 4), marking, and grading (as shown in Figure 5), all facilitated by the collaboration of multiple AI models like Gemini and ChatGPT. We use pre-trained models for their efficiency, given that the number of essays we can gather in Hong Kong is considerably smaller than the extensive labelled and unlabelled data utilized by large-scale pre-trained AI models. Han et al. (2021) underscore the advantage of using pre-trained AI models as a foundation rather than developing a new model from scratch.

* (一) 作文原文：

--- OCR Start ---
敬啟者：

本人為太子道花園大廈住客，近日大廈門口出現許多無牌小販擺賣，情況日趨嚴重，本人深感困擾，特此致函投訴。

首先，無牌小販霸佔大廈門口行人通道，阻礙居民出入。在上下班繁忙時間，市民出入受阻，需繞道而行，十分不便。

其次，他們整天大聲叫賣，十分嘈吵。有些學生需要溫習，他們的叫喊聲令學生無法專心，亦妨礙大廈居民休息。

此外，自從無牌小販出現後，街道變得十分骯髒。人們棄置食物殘渣和垃圾，吸引和滋生蚊蟲、老鼠等，令衛生環境惡化，嚴重影響社區整潔。

大廈附近的無牌小販日益猖獗，嚴重影響居民生活，本人促請大廈物業管理處迅速處理此事，給居民一個交代，還我們一個乾淨整潔的居住環境。如有任何問題，請致電 5364798 。

此致

太子道花園大廈物業管理處負責人 住客

2024 年5 月20 日

投訴人趙思懿敬上
--- OCR End ---

Fig.4 A random sample of OCR by the system

三、作文三評分表

評分項目	最高分	實際得分	評分依據
內容	16	15	清晰表明寫信原因 (2/2)；列舉三項投訴原因，描述清晰且具體，每項給予 5分，共計 15 分 (15/12)。
格式	15	12	啟首語 (3/3)；結束語及署名「投訴人」正確，但「台鑒」位置錯誤，扣 3分 (6/9)；日期格式正確 (3/3)。
標點字體	3	3	標點符號使用正確，段落劃分合理，字體工整。
錯別字	-3	-1	電話號碼缺失數字，扣 1分。
總分	34	29	

內容分析及修改建議 (三篇作文共通)

三篇作文都清楚地表達了投訴的理由，然而，可以透過加入更生動的例子、更精確的數據，以及更強而有力的詞語來提升文章的說服力。例如，可以將「十分吵雜」改為「叫賣聲此起彼伏，音量巨大，嚴重影響居民睡眠」，或加入具體數據，例如「小販佔據了約五米長的通道」。此外，可以更具體地說明噪音和衛生問題對居民的影響，例如，影響睡眠質素，引致精神不佳，影響工作效率，甚或影響身心健康。

格式分析及修改建議 (三篇作文共通)

大部分同學都能掌握投訴信的基本格式，但部分細節需要改進，例如「敬啟」或「台鑒」的位置應在收件人姓名之下，而非並列。 此外，投訴人的姓名應寫完整，避免造成聯絡上的不便。

標點及錯別字

同學們應仔細檢查標點符號的使用，並避免出現錯別字，確保文字表達準確無誤。

總結

三篇作文各有優劣，同學們已初步掌握投訴信的寫作技巧，但仍有進步空間。希望同學們在日後的寫作中，多加練習，注意細節，力求做到內容充實、邏輯清晰、格式正確、文字精煉，寫出更具說服力的投訴信。 特別是作文三，其內容描述較為詳盡，同學應繼續保持此良好的寫作習慣。

Fig.5 Random samples of marking reports by the system

The initial testing of the system involved evaluating a sample set of essays to assess grading accuracy and feedback quality. The results indicate that the current AI model and its configurations require further refinement. The system frequently struggles to recognize certain words in various compositions accurately or delivers incorrect corrections. Other researchers have identified similar issues. Several studies indicate that ChatGPT alone achieves a disappointing accuracy rate, ranging from 51% to 57% (Mizumoto & Eguchi, 2023; Kim et al., 2024). Furthermore, some researchers contend that AI may not yet possess the sophistication required to evaluate complex writing elements, apply scoring criteria effectively, or consider emotional nuances as adeptly as humans (Bui & Barrot, 2024). As a result, we are currently conducting tests to assess the accuracy of integrating various pre-trained AI models to enhance reliability.

In the second phase, we will concentrate on fine-tuning the AI model with more intricate prompts and incorporating the capability to upload scoring samples marked by humans for reference. This will enable the AI to utilize these samples when evaluating students' essays, and is the most effective for ensuring that the AI aligns closely with human evaluations. Lastly, we intend to collect sample essays from local primary and secondary schools and invite teachers to provide feedback to facilitate further improvements. The system's validity and reliability will be thoroughly assessed and customized for the context of Hong Kong before its official release.

5. Conclusion

The AI-empowered Chinese Composition Assessment System significantly advances automated essay evaluation, particularly in Chinese writing assessment. Our pilot testing indicates that the system can substantially reduce grading workloads for educators while providing timely and constructive feedback to students, which is crucial for fostering an engaging learning environment. By employing a multi-model grading approach, we have enhanced the reliability and validity of evaluations, ensuring fair and accurate assessments. While the system shows promise, further refinements are essential to fully capture the intricacies of human writing and ensure that assessments truly reflect student capabilities.

We have integrated existing LLMs like ChatGPT, Gemini, and Llama into a MaaS framework. Our user-friendly prompt framework allows educators to customize assessment criteria, aligning automated grading with human evaluations for more relevant feedback. As the landscape of AES in Chinese writing evolves, our work addresses a notable gap in the literature. We aim to enhance the sophistication and reliability of essay evaluations by optimizing

prompting techniques that better reflect human assessment, which could ultimately inform best practices in writing instruction.

In conclusion, while the AI-empowered Chinese Composition Assessment System shows great potential, it also underscores the complexities of automating writing assessment. Future research will focus on refining the system with diverse human-scored samples, ultimately striving to create an assessment tool that not only streamlines grading but also enriches the educational experience for educators and students.

Acknowledgements

The project was supported by the President's Development Fund, EdUHK. Ethical approval was granted by the Human Research Ethics Committee of The Education University of Hong Kong with the reference number [2023-2024-0451].

References

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. arXiv preprint arXiv:1606.04289.

Attali, Y. & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *Journal of Technology, Learning, and Assessment,* 4(3).

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies*, 1-18.

Chan, K. K. Y., Bond, T., & Yan, Z. (2023). Application of an automated essay scoring engine to English writing assessment using many-facet rasch measurement. *Language Testing*, *40*(1), 61-85.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1*(2), 939-944.

Genlott, A. A., & Grönlund, Å. (2013). Improving literacy skills through learning reading by writing: The iWTR method presented and tested. *Computers & education*, *67*, 98-104.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., ... & Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225-250.

Ke, Z., & Ng, V. (2019). Automated Essay Scoring: A Survey of the State of the Art. In *IJCAI* (Vol. 19, pp. 6300-6308).

Kim, H., Baghestani, S., Yin, S., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. *Exploring artificial intelligence in applied linguistics*, 73-95.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210.

Lee, C., Wong, K. C., Cheung, W. K., & Lee, F. S. (2009). Web-based essay critiquing system and EFL students' writing: A quantitative and qualitative investigation. *Computer Assisted Language Learning*, *22*(1), 57-72.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5), 238-243.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3), 2495-2527.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, *4*(4)