《外國語文研究》第四十二期
2025 年 12 月 頁 43-72
DOI:10.30404/FLS.202512_(42).0003

生成性 AI 在寫作評估中的研究綜述：機遇、挑戰與研究展望

楊潔* 何璿子**

香港城市大學

摘要

自 2022 年 11 月 ChatGPT 發佈以來，生成式人工智慧（GenAI）的快速發展已在語言教學與學習的各個層面，特別是在寫作評估方面，引發了巨大的變革。現有研究已探討將 GenAI 工具應用於寫作評估中的可能性，但其結論仍存在一定爭議，這在一定程度上限制了我們對 GenAI 工具在寫作評估中，尤其是在寫作評分以及提供寫作回饋方面之有效性的理解。因此，有必要對現有研究進行整合性綜述，以進一步釐清 GenAI 工具在寫作評估中的角色與成效。本綜述採用主題分析法（Braun & Clarke, 2006），整合分析了 2022 年至 2024 年間的 18 項實證研究，並歸納出三個主要主題，包括：(a) GenAI 工具在作文評分中的有效性；(b) GenAI 工具在提供書面回饋方面的能力；(c) GenAI 工具的侷限性與所引發的顧慮。最後，本綜述亦探討未來 GenAI 工具（例如 ChatGPT）在協助教師進行寫作評分與提供寫作回饋方面的潛在可能性，並提出 GenAI 於寫作評估領域中未來可行的研

* 楊潔，香港城市大學英文系博士生

** 何璿子，香港城市大學英文系研究助理

究方向與教學啟示。

關鍵詞：生成性 AI、自動化寫作評量、自動化寫作評分、形成性評量

A Synthesis of Research on Generative AI in Automated Writing Evaluation: Opportunities, Challenges, and Future Directions

Jie YANG* Xuanzi HE**

City University of Hong Kong

Abstract

Since the release of ChatGPT in November 2022, the rapid development of generative artificial intelligence (GenAI) has transformed various aspects of language teaching and learning, particularly in automated writing evaluation. Numerous studies have explored the potential of integrating GenAI tools into automated writing evaluation; however, their findings have been somewhat inconsistent. This variability makes it difficult for us to understand GenAI's capabilities in essay scoring and feedback provision within language education. Consequently, a comprehensive and up-to-date synthesis of the existing literature is essential to clarify the role and effectiveness of GenAI in this context.

Using thematic analysis (Braun & Clarke, 2006), this synthesis reviewed 18 empirical studies conducted between 2022 and 2024, identifying three key themes: (a) the effectiveness of GenAI tools in essay scoring, (b) the capabilities of GenAI tools in providing written feedback, and (c) the limitations and concerns associated with their use. This review evaluates the potential applications of GenAI tools, such as ChatGPT, in supporting

* Jie YANG, PhD candidate of the Department of English, City University of Hong Kong

**Xuanzi HE, Research Assistant of the Department of English, City University of Hong Kong

teachers with automated essay scoring and feedback delivery. Additionally, it highlights emerging research directions and pedagogical implications for integrating GenAI into automated writing evaluation practices.

Key words: Generative AI, automated writing evaluation, automated essay scoring, formative feedback

1. Introduction

Generative artificial intelligence (GenAI) has attracted global attention since the release of ChatGPT in November 2022. Utilizing a machine learning (ML) model, GenAI demonstrated the ability to analyze large datasets, identify patterns, and produce coherent and grammatically correct text with ongoing improvements from user interactions (IBM 2024). These multiple affordances of GenAI tools exhibit the significant potential to revolutionize the field of language teaching, learning, and assessment (Hong, 2023). One promising area in language education is the integration of GenAI in Automated Writing Evaluation (AWE), which utilizes artificial intelligence and natural language processing techniques to assess and provide feedback on written text. A number of attempts have been made to investigate the use of GenAI in automated essay scoring (e.g., Bui & Barrot 2024; Mizumoto & Eguchi 2023) and providing formative feedback (e.g., Lu et al. 2024; Teng 2024). However, research findings yielded exhibit variability and a consensus on the effectiveness of GenAI tools in automated writing evaluation has not yet been reached.

Despite calls for evaluating the use of GenAI tools in automated writing evaluation (e.g., Burstein 2023), existing literature has not fully addressed how GenAI tools can be used. This is particularly important given the growing integration of technology in language teaching and learning. Thus, there is a critical need to examine how GenAI tools can be used in automated writing evaluation. To address this gap, the current study undertakes a synthesis of the up-to-date studies starting from 2022 on the use of GenAI tools in writing evaluation by adopting thematic analysis (Braun & Clarke 2006). By grounding themes in empirical data, the inductive thematic analysis provides a detailed and evidence-based description of GenAI's use in automated writing evaluation. It is expected to contribute to language teaching and learning by identifying the key challenges and opportunities in integrating GenAI into automated writing evaluation practices based on the extant literature. Additionally, it enables us to propose evidence-based strategies for optimizing GenAI-human collaboration in language classroom settings. It also allows us to advance theoretical models for evaluating GenAI's role in fostering writing development. Grounded in empirical data, the present study synthesizes the findings on the use of GenAI in automated writing evaluation, with a

particular focus on essay scoring and feedback. This review begins with a broad examination of GenAI applications in automated writing evaluation. The paper then examines prior syntheses of studies on the use of GenAI in a broad context of language teaching and learning to justify the need for an up-to-date research synthesis on the use of GenAI in automated writing evaluation. Next, the methodology guiding this review is clearly outlined. Building on this foundation, the findings and discussion section presents and critically discusses the key themes emerging from the analysis. Finally, the review concludes by outlining practical implications for educators and proposing directions for future research in this evolving field. By doing this, the up-to-date synthesis can inform both instructors and researchers about how GenAI can benefit automated writing evaluations, leading to more effective use of GenAI tools in language teaching and learning.

2. Literature review

2.1 An overview of GenAI in automated writing evaluation

Using large language models (LLMs), GenAI tools can be trained on large datasets to create a deep learning neural network and create new textual and multimodal content. Examples of some widely used GenAI tools are OpenAI's ChatGPT, Anthropism's Claude, Google's Gemini, and Runaway's Gen-2, etc. Among these GenAI tools, ChatGPT has been frequently used in language teaching and learning, particularly in automated writing evaluation. Utilizing a large language model known as a generative pre-trained transformer (GPT), the ChatGPT platform has shown its capacity to generate human-like texts and respond to diverse textual prompts, inquiries, and interactive dialogue scenarios. These affordances endow GenAI tools (e.g., ChatGPT) with considerable potential in facilitating Automated Writing Evaluation (AWE), which involves utilizing computer technology to analyze and assess written texts (Chen & Cheng, 2008). The AWE systems are capable of delivering both summative evaluation and formative feedback (Ranalli et al., 2017; Stevenson, 2016). When the focus is solely on summative evaluation, the term Automated Essay Scoring (AES) is often used interchangeably with AWE (Shermis & Burstein, 2013). These

summative evaluations are designed to offer an objective judgment of writing, usually serving as a complement to human-provided evaluation. For formative feedback, the feedback generated by AWE tools can range from simple corrections, such as grammar and spelling, to more advanced insights, such as recommendations for enhancing text coherence and structure (Stevenson, 2016). Therefore, following this line of definition, the GenAI-assisted AWE in the present review includes both formative feedback and summative evaluation (or in other words, automated essay scoring).

Several studies have been conducted to investigate the effectiveness of GenAI tools in automated essay scoring. (e.g., Bui & Barrot 2024; Li et al. 2024; Parker et al. 2023, etc.). These studies mainly followed the Technological Pedagogical Content Knowledge (TPACK) Framework, which includes the knowledge required to teach specific subjects and how technology can be used to facilitate essay scoring (Mishra & Koehler 2006). This framework provides a theoretical lens to analyze how teachers integrate GenAI tools into their writing evaluation practice, assess their effectiveness, and address challenges. By examining studies under the TPACK framework, we are able to gain insights into the practical implications of using GenAI tools in authentic writing evaluation contexts. However, consensus on the reliability of GenAI in automated essay scoring has not yet been achieved due to the mixed and sometimes discrepant research findings. The majority of studies (e.g., Li et al. 2024; Parker et al. 2023; Tate et al. 2024) demonstrated general agreement between the scores assigned by GenAI and human raters. However, a few exceptions (e.g., Bui & Barrot 2024; Shabara et al. 2024; Yancey et al. 2023) reported low or varying alignment of the scores provided by GenAI and human raters, indicating the limitations of GenAI in automated writing evaluation. Another group of studies examined the use of GenAI tools in providing formative feedback by analyzing the assessment data (i.e., corrective feedback) provided by GenAI following the Learning-Oriented Assessment (LOA) framework (Purpura, 2024; Turner & Purpura, 2016). Perceptions and practice of the use of GenAI and the impact of GenAI-provided feedback on students' writing motivation and engagement (e.g., Steiss et al. 2024; Teng 2024). Overall, a general consensus has been reached that GenAI can potentially supplement teacher feedback despite the inaccuracy and redundancy of the feedback provided.

2.2 Past reviews of GenAI in language teaching and learning

While a group of studies have investigated the use of GenAI in automated writing evaluation, to the best of my knowledge, few have systematically synthesized these findings to provide a comprehensive understanding of its effectiveness, limiting the ability to inform pedagogical practices and future research. Instead, existing syntheses primarily examined the use of GenAI in a broader domain of language teaching and learning (e.g., Law 2024; Li et al. 2024). For example, Law (2024) reviewed 41 papers published between 2017 and July 2023 on the use of GenAI in language teaching and learning. While acknowledging the potential of GenAI tools in writing evaluation, the author suggested that a comprehensive understanding of the effectiveness of GenAI in diverse educational settings is needed. Similarly, Li et al. (2024) probed into the use of ChatGPT in language teaching and learning by critically reviewing 36 articles between November 2022 and November 2023. This article mentioned the role of ChatGPT in correcting vocabulary and grammar in students' writing to achieve higher writing quality. The authors further proposed that ongoing studies are essential to validate the effectiveness of ChatGPT across various contexts and determine optimal strategies for its implementation. These reviews provide valuable insight into the current research trends and future directions in the use of GenAI tools in language teaching and learning.

Despite the valuable contributions of these previous reviews, considering the rapid advancements of GenAI tools, there is still a need for an up-to-date and comprehensive review of the studies on the application of GenAI in automated writing evaluation, particularly in automated essay scoring and formative feedback. When evaluating an essay, AWE systems adopt various methods such as natural language processing (NLP), machine learning (ML), and artificial intelligence (AI), as well as advanced statistical methods (Grimes & Warschauer, 2010). Some widely investigated AES systems are Criterion, My Access, Grammarly, Writing Mentor, etc. However, the focus of the present review is the use of GenAI tools in AWE because the release of ChatGPT in November 2022 provides innovative approaches to AWE. Different from prior AI tools, GenAI tools can be trained on large datasets using large language models (LLMs) to

create a deep learning neural network and create new textual and multimodal content.

Therefore, the present review covers studies in 2022-2024 due to the new but rapidly evolved GenAI tools and the proliferation of studies during 2023-2024 that specifically investigate the use of GenAI in automated essay scoring and formative feedback. By reviewing recent studies, we are able to identify the research gaps and emerging trends, thereby guiding future research endeavors in this field. Furthermore, the findings derived from the review have practical implications as they can support evidence-based decision-making in automated writing evaluation under the assistance of GenAI. The present research synthesis aims to provide an up-to-date synthesis of the literature, focusing on empirical studies that have examined the use of GenAI tools (predominantly Chat GPT, plus some emerging tools such as EvaluMate) in automated writing evaluations across various contexts, and to critically evaluate the potential of GenAI tools in revolutionizing practices in automated writing evaluations.

3. Method

3.1 Data set

The selection procedure for the studies to be included in this research synthesis adheres to the PRISMA principles (Moher et al. 2009), as shown in Figure 1. Studies on the use of GenAI in automated writing evaluation, including automated essay scoring and formative feedback were included. These studies were selected from Scopus and Web of Science due to their wide coverage of the extensive indexing of esteemed academic journals in the field of applied linguistics and language teaching and learning. Given that GenAI in automated writing evaluation is an emerging area and research has been conducted since ChatGPT's release on November 30, 2022, the search period for this synthesis was limited to 2022-2024. Eligible publication types include published empirical journal articles and conference papers due to their significance in capturing the latest evidence and shaping research direction at the early-stage GenAI technological advancements. Potential literature was identified using the following search terms:

generative artificial intelligence OR generative AI OR GenAI OR ChatGPT

OR Chat GPT AND writing evaluation OR writing assessment OR feedback OR scoring

Additionally, to comprehensively include all relevant literature, a hand search via Google Scholar of journals in the field of language education and technology was conducted.

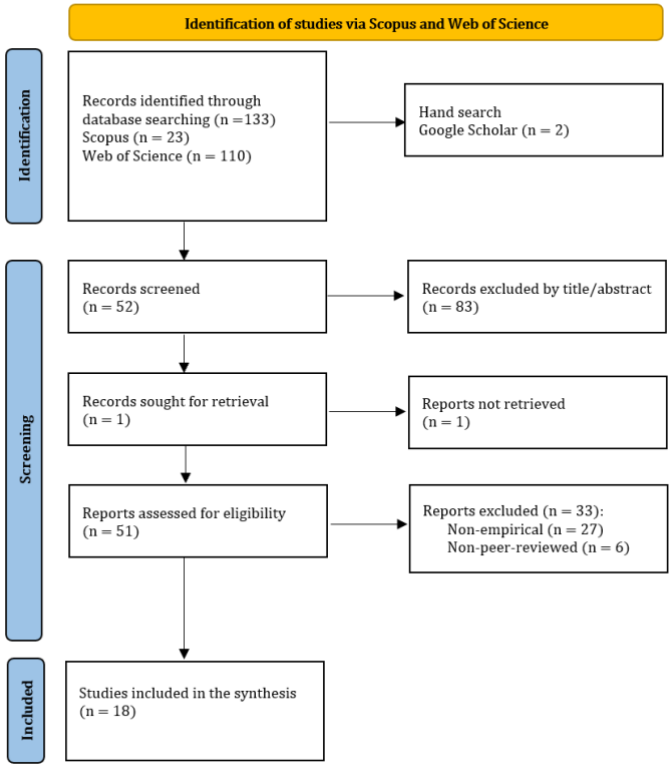

Figure 1. PRISMA flow diagram representing the data selection process

3.2 Criteria for exclusion

The preliminary search yielded 135 papers. To identify the research relevant to the synthesis, several exclusion criteria were considered in the screening phase. The first criterion was the relevance of the research scope. The two researchers carefully read the titles, abstracts, and research questions of the studies and excluded the articles that did not address the use of GenAI in automated writing evaluation (n=83). Then, the article that was not retrieved due to limited access to the journals was excluded (n=1). The second criterion was that the identified studies should be empirical in nature. Thus, non-empirical studies such as reviews and commentaries were excluded (n=27). Last, to ensure the quality and reliability of the synthesis, only articles published in peer-reviewed journals were included to ensure the quality of the research. Therefore, non-peer-reviewed articles were excluded (n=6). For the conference paper, we browsed the official website of the conference, and according to the official guidelines, all conference papers were peer-reviewed. Furthermore, this study has been cited over 80 times according to Google Scholar, which, to some extent, indicates the reliability of the study. Therefore, the conference paper was included in the synthesis. After a comprehensive review, a total of 18 articles were finalized for this synthesis.

3.3 Data analysis

In this research synthesis, thematic analysis (Braun & Clarke 2006) was employed to identify, analyze, and report patterns within the data related to the use of GenAI in automated writing evaluation. Specifically, the inductive thematic analysis was employed to identify the emerging themes in terms of the use of GenAI in automated writing evaluation. The inductive thematic analysis is a process of coding the data without trying to fit it into a pre-existing coding frame or the researcher's analytic preconceptions, thus the themes identified are closely related to the data themselves (Patton 1990). In the present review, we did not prepare any pre-designed research questions or refer to any pre-existing coding frame. Instead, we followed the procedure of thematic analysis and formed the themes based on the content of the empirical studies (e.g., research questions, research contexts, research findings, limitations, etc.)

To ensure reliability and transparency in the thematic analysis, two researchers with experience in language teaching and learning were

involved throughout the process. Following Braun and Clarke's (2006) step-by-step guide, we started by familiarizing ourselves with the data by reading and rereading the research articles and taking notes of the initial ideas. This was followed by generating initial codes that captured the essence of the data, which were then organized into potential themes. Specifically, we copied and pasted the content about the use of GenAI in AWE and color-coded the texts to indicate the extent to which the content is positive/negative/neutral. After that, we extracted the preliminary themes from the color-coded texts and summarized these themes in phrases and short sentences. These themes were continuously refined through an iterative process of coding and recording to ensure that they accurately reflected the most salient aspects of the data. In this process, both researchers independently reviewed and coded the data to identify initial themes. Afterward, we compared our codes and discussed any differences to reach a consensus, ensuring consistency in the analysis (Cohen's $\kappa = 0.82$, indicating strong agreement). This collaborative approach helped minimize individual bias and strengthened the reliability of the findings. The next step involved defining and naming the themes by going back to the data for each theme and organizing them in a coherent and consistent way. Additionally, an audit trail was maintained to document all coding decisions and revisions, providing a clear record of how the analysis progressed. To further enhance transparency, the themes were also shared with a group of participants for feedback, ensuring the findings accurately resonate with their experiences. These steps ensured that the thematic analysis was both rigorous and trustworthy. Finally, based on a set of fully developed themes, we presented the findings with sufficient evidence within the data to show how GenAI has been used in automated writing evaluations. Table 1 provides an overview of the themes, sub-themes, and sample studies in this synthesis.

Table 1: Overview of themes and sub-themes that emerged in the research synthesis

Themes	Subthemes [number of studies evidencing each subtheme]	Sample studies
Effectiveness of GenAI tools in essay scoring	Accuracy of GenAI scoring [11]	Geçkin et al. 2023; Mizumoto & Eguchi 2023
	Consistency of GenAI scoring [4]	Bui & Barrot 2024; Tate et al. 2024
Capabilities of GenAI tools in providing written feedback	Comparison with teacher-provided feedback [7]	Li et al. 2024; Guo & Wang 2024
	Impact on students' writing practice and performance [3]	Guo et al. 2024; Teng 2024
	Students' and teachers' perceptions of GenAI- provided feedback [3]	Teng 2024; Guo & Wang 2024
Limitations and concerns elicited	Issues with research design [16]	Shin & Lee 2024; Guo et al. 2024
	Limitations and ethical concerns about the use of GenAI tools [14]	Shabara et al. 2024; Su et al.,2023;

This synthesis reviewed 18 articles exploring GenAI in automated writing evaluation (see Table 2), focusing on automated essay scoring (AES) (eight studies), feedback (seven studies), and both (three studies). Among these studies, the quantitative approach (nine studies) and mixed approach (eight studies) were the most widely used, with one study adopting the qualitative approach. For genres being investigated, the argumentative essay was the most frequently explored (ten studies). The majority of studies were conducted in the EFL/ESL context (eleven studies) and in the college settings (ten studies), with less attention being paid to secondary-level settings (three studies).

Table 2: Key information from sample studies

Study	Focus	Type	GenAI tool	Genre	Setting	Language context
Geçkin et al. 2023	AES	Quantitative	ChatGPT-3.5	Paragraph writing task	College	EFL
Bui and Barrot 2024	AES	Quantitative	ChatGPT-3.5	Argumentative essay	College	ESL
Guo and Wang 2024	Feedback	Mixed	ChatGPT (version not specified)	Argumentative essay	College	EFL
Guo et al. 2024	Feedback	Quantitative	EvaluMate	Argumentative essay	College	EFL
Li et al. 2024	AES and feedback	Mixed	ChatGPT-3.5 and ChatGPT-4	Argumentative essay	College	EFL
Lin and Crosthwaite 2024	Feedback	Mixed	ChatGPT-4	Argumentative essay	Mixed	EFL&ESL
Lu et al. 2024	AES and feedback	Mixed	ChatGPT-3.5	Article abstract	College	Chinese as a Native Language
Mizumoto and Eguchi	AES	Quantitative	ChatGPT-3.5	Essays	Not specified	EFL

2023						
Parker et al. 2023	AES and feedback	Mixed	ChatGPT-3	Mixed genres	Undergraduate and graduate	Not specified
Shabara et al. 2024	AES	Quantitative	ChatGPT-3.5	Expository	College	EFL
Shin and Lee 2024	AES	Quantitative	ChatGPT-4 (customized chatbot)	Argumentative essay	Secondary level	EFL
Su et al. 2023	Feedback	Qualitative	ChatGPT (version not specified)	Argumentative essay	Not specified	Not specified
Steiss et al. 2024	Feedback	Mixed	ChatGPT-3.5	Argumentative essay	Secondary school	Mixed
Tate et al. 2024	AES	Quantitative	ChatGPT-3.5 and ChatGPT-4	ELA (English language arts) and history	Secondary level	Mixed
Teng 2024	Feedback	Mixed	ChatGPT (version not specified)	N.A.	College	EFL
Wang et al. 2024	Feedback	Mixed	ChatGPT (version not	Argumentative essay	College	Chinese as a Native Language

			specified)			
Yamashita 2024	AES	Quantitative	ChatGPT-4.0	Argumentative essay	Not specified	ESL
Yancey et al. 2023	AES	Quantitative	ChatGPT-3.5 and ChatGPT-4	Short essay response	Not specified	Mixed

4. Findings and discussion

Three themes on the use of GenAI in automated writing evaluation were identified, including (a) effectiveness of GenAI tools in automated essay scoring, (b) capabilities of GenAI tools in providing written feedback, and (c) limitations and concerns elicited.

4.1 Effectiveness of GenAI tools in automated essay scoring

In the reviewed literature, the effectiveness of GenAI on automated writing evaluation is predominantly assessed by the accuracy (mainly evaluated by the agreement with human-assigned scores) and internal consistency (mainly evaluated by the agreement with scores assigned by GenAI over multiple iterations). This synthesis identified a general agreement in the supportive and supplementary role of GenAI tools in essay scoring (evident in eight studies), despite the discrepancies in the degree of accuracy and consistency reported across studies. For instance, Shin and Lee (2024) built a customized chatbot based on ChatGPT 4 and then compared the scores of the 50 English essays written by Korean EFL students assigned by the chatbot and two English teachers. The results of the correlation analysis indicated a strong similarity between the scores given by the chatbot and English teachers. This study highlighted the role of GenAI in providing accessible and valuable support for human evaluation, especially for instructors with limited English proficiency and training on rating.

Furthermore, the improved effectiveness, represented by higher accuracy and consistency of GenAI was demonstrated along with the GenAI technology advancement. This overall pattern is reflected in several studies. For example, four studies (e.g., Bui & Barrot 2024; Geçkin et al. 2023; Parker et al. 2023; Shabara et al. 2024; Yancey et al., 2023) explored the effectiveness of ChatGPT 3.0/3.5 in essay scoring and reported low agreement as well as weak to moderate correlations between scores assigned by ChatGPT and human raters. Additionally, the lower scores with greater deviation imply a limited capacity of GenAI in essay scoring. An example is Bui and Barrot (2024), which examined the relationship between the scores given by ChatGPT and an experienced rater by comparing the scores of 200 argumentative essays written by English L2 learners from Asia countries. Results revealed a weak to moderate correlation between the scores given by ChatGPT and the human rater, indicating a weak alignment. They also delve into the internal consistency of scores provided by ChatGPT over multiple iterations, and similarly, they found that the scores failed to establish consistency. As GenAI technology advances, more recent studies (e.g., Shin & Lee 2024; Tate et al. 2024; Yamashita 2024) employed ChatGPT 4 in automated essay scoring and yielded more promising results. For instance, Tate et al. (2024) reported better internal consistency of the scores assigned by ChatGPT 4 and a higher level of agreement (fair to moderate) between ChatGPT 4 and human-provided scores.

Although the performance of earlier versions of ChatGPT in automated essay scoring tends to be unsatisfactory, an interesting exception to the overall trend is observed in Lu et al. (2024), which warrants further exploration. In this study, the authors examined the writings of 46 Chinese undergraduate students in southern China for an academic writing task. By comparing the scores given by ChatGPT 3.5 and human raters, moderate to good agreement was yielded. However, it should be noted that the writings being assessed are Chinese article abstracts, which is different from the majority that assessed English argumentative essays. This indicates that the language used and the genre of writing could exert an impact on the effectiveness of GenAI tools.

4.2 Capabilities of GenAI tools in providing written feedback

For the capabilities of GenAI tools in giving feedback on writing, the great

potential and drawbacks have been acknowledged across studies, leading to a call for the integration and collaboration of GenAI and human instructors in generating feedback. This has been supported by three groups of evidence, including comparison with teacher-provided feedback, impact on students' writing performance, and students' and teachers' perceptions of GenAI-generated feedback.

Compared with teacher-produced feedback that features consistency, higher quality, and humanistic empathy interaction (e.g., Lin & Crosthwaite,2024; Steiss et al. 2024; Wang et al. 2024), GenAI tools are inclined to provide feedback with greater variance and give redundant feedback on local issues (e.g., sentence-level grammar). Additionally, as reported by Wang et al. (2024), GenAI tools are incapable of evaluating students' arguments when offering feedback and can only generate limited affective feedback (e.g., You're making progress with your writing!) at the linguistic level. Despite the observed limitations, the distinctive advantages of GenAI enable it to effectively complement human instructors in providing feedback. Several potential affordances of GenAI in giving feedback discussed in the reviewed studies are: (a) supports multiple submissions (e.g., Lu et al. 2024; Parker et al. 2023), (b) provides more comprehensive and equal feedback across multiple dimensions (e.g., Guo & Wang 2024; Li et al. 2024), (c) offers dynamic, fluent, and multi-turn dialogue feedback with great ease (e.g., Steiss et al. 2024; Wang et al. 2024).

Moreover, this synthesis examined the capabilities of GenAI in terms of the impact on students' writing practice and performance, and an overall positive influence was identified (evident in three studies). First, GenAI-provided feedback can enhance students' writing motivation and understanding of teachers' assessments, facilitating independent thinking about their writing (e.g., Lu et al. 2024; Teng 2024). For instance, adopting a mixed-method research design, Teng (2024) investigate EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. By analyzing the quantitative data from the questionnaire and the qualitative data from the interview with 45 EFL learners in Macau, the author found positive effects of the use of GenAI tools on writing motivation, engagement, self-efficacy, and collaborative writing tendency. Second, the use of GenAI in producing feedback contributes to greater improvements

in students' writing performance and feedback quality (e.g., Guo et al. 2024). Third, GenAI tools, such as ChatGPT, can keep the conversation history with students. These detailed records enable learners to track their writing process and monitor progression, fostering self-regulated learning and iterative revision, which are critical to improving students' writing practice (e.g., Su et al. 2023).

Furthermore, two studies (Guo & Wang 2024; Teng 2024) reviewed in the synthesis concluded that learners and instructors expressed largely positive perceptions of the use of GenAI in feedback. Specifically, the majority of students appreciated the accuracy, reliability, and convenience of GenAI in producing feedback. From teachers' perspective, the use of GenAI can lessen their feedback burden, reduce their workload, and improve their feedback literacy by enabling them to focus on multiple aspects of students' writing. However, several limitations, including providing lengthy, irrelevant, and inappropriate feedback, as well as the inequity caused by the inaccessibility of ChatGPT in some areas, were also acknowledged (e.g., Guo & Wang 2024). Although a complete picture of people's perceptions of GenAI in providing feedback is yet challenging to achieve due to the limited studies, it is reasonable to assume that GenAI holds a place in supplementing teachers' feedback practice.

4.3 Limitations and concerns elicited

The limitations and concerns identified in the synthesis involve two aspects: issues with research design and limitations and ethical concerns in the use of the GenAI tools. For research design, the limited methodological rigor and scope of analysis constrained a comprehensive understanding of the use of GenAI in automated writing evaluation. First, the small sample sizes, writing genres, and prompts in the studies may impede broader conclusions. Second, the lack of control for potential confounding variables, such as assessment rubrics and human bias, can result in overgeneralization of the research findings. In terms of the scope of analysis, existing research examined restricted feedback types and error types (e.g., Guo et al. 2024; Lin & Crosthwaite,2024; Teng 2024), potentially constraining our comprehension of the topic.

This synthesis also identified several limitations of the GenAI tools and ethical concerns that emerged from the reviewed studies (e.g., Bui & Barrot 2024; Mizumoto & Eguchi 2023). First, the black-box-like nature of

the deep learning models resulted in low interpretability and explicitness of the process of GenAI-assisted automated writing evaluation. Additionally, GenAI tools are incapable of capturing the multidimensionality of writing and providing human-like emotional interaction. Third, the logical reasoning deficiency of GenAI tools may lead to irrelevant, over-abstract, lengthy, or fake responses (identified in four studies). For instance, Teng (2024) noted that some GenAI-provided feedback sounds too formal, lacks a personal touch, and is hard to follow. GenAI also sometimes offers off-topic comments and struggles to identify specific writing issues. Although the limitations of GenAI may impede its use in wider contexts, they can potentially be addressed by rapid technological advancements.

For ethical concerns, five studies expressed concerns related to authorship, plagiarism, confidentiality, potential bias, overreliance and dehumanization of educational practices (e.g., Mizumoto & Eguchi 2023; Parker et al. 2023; Su et al. 2023), highlighting the need for responsible and principled implementation of GenAI in writing classrooms. To support the responsible use of GenAI, these concerns should be thoughtfully considered. In terms of authorship and plagiarism issues, the vague boundaries of human-AI collaboration in text generation raise questions about intellectual ownership. To address this, institutions could adopt explicit guidelines (e.g., using the CRediT taxonomy to acknowledge AI contributions) and integrate AI-specific citation protocols into academic integrity policies. For course instructors, they can customize the regulations and policies of the use of GenAI in assignments according to the Intended Learning Outcomes (ILOs) of the course. In terms of confidentiality, GenAI platforms often require data input that may compromise users' privacy. Possible strategies to mitigate this issue are prioritizing tools with transparent data protection policies (e.g., GDPR compliance) and educating users about data anonymization practices to raise their awareness of data protection. In terms of concerns about overreliance and dehumanization, AI should play a complementary rather than substitutive role in automated writing evaluation. For example, AI-generated feedback could be paired with discussion with peers or instructors to preserve interpersonal engagement.

In conclusion, the synthesis highlights key limitations and concerns, providing crucial guidance for future research. Addressing these issues will not only deepen our knowledge of GenAI's role in automated writing evaluation but also inform pedagogical practices, leading to more effective integration of GenAI tools in writing instruction. Moreover, considering the ethical concerns while using GenAI tools cautiously in writing classrooms is also necessary for more responsible and effective integration into educational practices. Future research could explore institutional partnerships to develop standardized ethical frameworks for AI adoption in education.

5. Pedagogical implications and conclusion

In an attempt to identify the opportunities, challenges and future directions in the use of GenAI in automated writing evaluation, this research synthesis reviewed 18 papers published in the past three years and identified several key themes. These reviewed studies predominantly explored the effectiveness of GenAI tools in automated essay scoring, the capabilities of GenAI tools in providing written feedback, and limitations and concerns elicited, resonating with the trend of practice-based research on L2 writing identified by Sun and Lan (2023).

The findings provide useful insights for pedagogy to facilitate more effective integration of GenAI in automated writing evaluation. On the one hand, the innovative GenAI tools provide great opportunities for assisting and facilitating automated writing evaluations. This is supported by the overall positive results in the effectiveness and people's perceptions of the use of GenAI in automated writing evaluation identified in the synthesis. In particular, for automated essay scoring, the latest GenAI tools (e.g., ChatGPT 4o and Deepseek) demonstrated general satisfactory accuracy and consistency, implying that GenAI tools can be sufficient for low-stakes and formative assessment, which can alleviate teachers' burden in marking essays. For providing feedback, GenAI tools are capable of providing instant and textual-based feedback across language, content, and organization following a well-designed prompt. The supportive role of GenAI in providing feedback has also been reflected in the notable positive effect on students' motivation, as the accuracy, reliability, and ease of using GenAI in generating feedback has been fully acknowledged by students. These strengths make GenAI tools a valuable supplement to teachers'

feedback, particularly in formative early drafts.

On the other hand, the drawbacks of the GenAI tools and people's limited understanding of GenAI tools pose significant challenges to their utilization in writing classrooms. First, since it has been proved that GenAI tools exhibit varied responses depending on the prompts provided, one challenge users need to address is to generate high-quality prompts to improve the performance and ensure the accuracy of GenAI tools in automated essay scoring and feedback. Second, due to the potential limitations of GenAI tools in producing inaccurate, lengthy, and irrelevant feedback, it is important for students and teachers to equip themselves with critical AI literacy to make judgments upon the responses. Third, the potential challenges in privacy, potential bias, and academic integrity need to be considered to ensure responsible and equitable implications of GenAI in language education.

To address the above-mentioned ethical and pedagogical challenges and ensure effective and responsible integration of AI into education, it is crucial to enhance instructors' AI literacy (Hur 2025). This involves: 1) developing a foundational understanding of AI systems and the skills to communicate and collaborate effectively with them (Long 2020; Allen 2023), 2) fostering the capacity to critically reflect on AI applications in teaching (Ding 2024) and adaptively integrate them into pedagogical contexts (Ng et al. 2021), and 3) cultivating the ability to evaluate AI tools in terms of fairness, accountability (Salhab 2024), transparency, safety, ethical considerations, and their broader societal impact (Kočková 2024). One promising approach to improving instructors' AI literacy is through targeted professional development and training programs. For example, institutions could adopt training frameworks such as UNESCO's AI Competency Framework for Teachers (2022)*, which emphasizes ethical AI use, data privacy, and bias mitigation. Based on the framework, workshops can be held to provide training on theoretical foundations (e.g., the algorithmic bias) and hands-on practice (e.g., designing AI-augmented writing tasks). Additionally, open-access resources can also be used to

* https://unesdoc.unesco.org/ark:/48223/pf0000391104

enhance instructors' AI literacy. Platforms like Coursera[**]and ISTE AI[***] offer free courses on AI tool evaluation and integration. Peer-learning platforms such as EducateAI[****] could foster knowledge-sharing on AI tool efficacy and classroom adaptation, and instructors could co-design rubrics to assess student-AI collaboration or share anonymized case studies on mitigating overreliance on AI tools. To sum up, by making good use of the training programs and open-access sources, instructors can improve their AI literacy and transition from passive consumers to critical co-designers of AI-assisted pedagogy.

For the future research agenda, the reviewed research highlighted the role of prompts in the use of GenAI due to its key affordance in interacting with users. However, little has been known about how to prompt GenAI tools for the best performance in automated essay scoring and feedback. Therefore, future studies may place more emphasis on prompt design and examine diverse prompting methods in various situations to gain more insights into how GenAI tools can be trained for improved performance in automated writing evaluation. Furthermore, additional research could dig deeper into how GenAI tools modify their responses and cope with different requests (e.g., to evaluate multimodal writings or different writing genres). This resonates with the emphasis on prompt programming in earlier studies (e.g., Reynolds & McDonell 2021). Moreover, as Davis (1989) noted, the effective implementation of any technology depends on a deeper comprehension of the user acceptance processes. Therefore, in response to Parker et al.'s (2023) call for more research attention on user acceptance of GenAI, future studies could explore students' and teachers' acceptance of the GenAI tools in automated writing evaluation. Moreover, considering the crucial role of cultural background and people's way of thinking in determining people's acceptance of technological innovation (Jan et al. 2022; Sun et al. 2019),

[**] https://www.coursera.org/

[***] https://iste.org/ai

[****] https://www.educate-ai.com/

exploring cross-cultural differences in the use of GenAI presents an important avenue for future research. Such explorations can be conducted in contexts like the USA and HK, where people may have diverse nationalities. Future investigations could explore how cultural norms, values, and communication styles influence the adoption, perception, and outcomes of GenAI tools in classrooms by collecting data through class observation and interviews. For instance, researchers might examine whether students from cultures with a higher tolerance for uncertainty are more receptive to AI-generated texts or how varying cultural attitudes toward creativity and automation influence GenAI's effectiveness in different contexts. These comparative studies across diverse cultural settings could provide valuable insights into tailoring GenAI systems to meet the specific needs and preferences of language instructors, researchers, and students around the world. Future studies may also explore the best practices in different cultural contexts to develop users' critical awareness of AI literacy and empower them to achieve effective and ethical use of GenAI in language teaching and learning

References

Adams, Thomas, et al. “Education for Sustainable Development: Mapping the SDGs to University Curricula.” *Sustainability*, vol. 15, no. 10, 20 May 2023, p. 8340, https://doi.org/10.3390/su15108340. Accessed 3 Sept. 2023.

Allen, Laura Kristen, and Panayiota Kendeou. "ED-AI Lit: An Interdisciplinary Framework for AI Literacy in Education." *Policy Insights from the Behavioral and Brain Sciences*, vol.11, no.1, 18 Dec. 2023, pp. 3-10, https://doi.org/10.1177/23727322231220339.

Barrot, Jessie S. “Using ChatGPT for Second Language Writing: Pitfalls and Potentials.” *Assessing Writing*, vol. 57, 25 May 2023, p. 100745, https://doi.org/10.1016/j.asw.2023.100745.

Braun, Virginia, and Victoria Clarke. “Using Thematic Analysis in Psychology.” *Qualitative Research in Psychology*, vol. 3, no. 2, 21 July 2008, pp. 77–101, https://doi.org/10.1191/1478088706qp063oa.

Bui, Ngoc My, and Jessie S. Barrot. “ChatGPT as an Automated Essay Scoring Tool in the Writing Classrooms: How It Compares with Human Scoring.” *Education and Information Technologies*, vol. 30, 13 July 2024, pp. 2041-58, https://doi.org/10.1007/s10639-024-12891-w.

Chapelle, Carol A., et al. “Paths for Exploring AI in Applied Linguistics.” *Exploring artificial intelligence in applied linguistics,* edited by Carol A. Chapelle, Gulbahar H. Beckett and Jim Ranalli. Iowa State University Digital Press, 31 July 2024, pp. 1-8, https://doi.org/10.31274/isudp.2024.154.01.

Chen, Chi-Fen Emily, and Wei-Yuan Eugene Cheng. "Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes." *Language Learning and Technology*, vol. 12, no. 2, June 2008, pp. 94–112.

Davis, Fred D. “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology.” *MIS Quarterly*, vol. 13, no. 3, Sept. 1989, pp. 319-40, https://doi.org/10.2307/249008.

Ding, Ai-Chu Elisha, et al. "Enhancing Teacher AI Literacy and Integration through Different Types of Cases in Teacher Professional

Development." *Computers and Education Open*, vol. 6, 10 Apr. 2024, pp. 100178, https://doi.org/10.1016/j.caeo.2024.100178.

Duman, Guler, et al. "Research Trends in Mobile Assisted Language Learning from 2000 to 2012." *ReCALL*, vol. 27, no. 2, 21 July 2014, pp. 197–216, https://doi.org/10.1017/s0958344014000287.

"GDPR Overview – How to Stay Compliant and Protect Data." Coppa Online Security & Control Organizations, 13 June 2024, www.coppa.org/gdpr. Accessed 18 Feb. 2025.

Geçkin, Vasfiye, et al. "Assessing Second-Language Academic Writing: AI vs. Human Raters." *Journal of Educational Technology and Online Learning*, vol. 6, no. 4, 31 Dec. 2023, pp. 1096–1108, https://doi.org/10.31681/jetol.1336599.

Grimes, Douglas, and Mark Warschauer. "Utility in a fallible tool: A multi-site case study of automated writing evaluation." The Journal of Technology, Learning and Assessment, vol. 8, no. 6, 02 Mar. 2010, https://ejournals.bc.edu/index.php/jtla/article/view/1625.

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal,* 26(2), 91-108.

Guo, Kai, and DeLiang Wang. "To Resist It or to Embrace It? Examining ChatGPT's Potential to Support Teacher Feedback in EFL Writing." *Education and Information Technologies*, vol. 29, 29 Aug. 2023, pp. 8435–63, https://doi.org/10.1007/s10639-023-12146-0.

Guo, Kai, et al. "Effects of an AI-Supported Approach to Peer Feedback on University EFL Students' Feedback Quality and Writing Ability." *The Internet and Higher Education*, vol. 63, 14 July 2024, p. 100962, https://doi.org/10.1016/j.iheduc.2024.100962. Accessed 22 Sept. 2024.

Hampton, Stephanie E., and John N. Parker. "Collaboration and productivity in scientific synthesis." *BioScience*, vol. 61, no. 11, 01 Nov 2011, pp. 900-10, https://doi.org/10.1525/bio.2011.61.11.9

Hong, Wilson Cheong Hin. "The Impact of ChatGPT on Foreign Language Teaching and Learning: Opportunities in Education and

Research." *Journal of Educational Technology and Innovation*, vol. 5, no. 1, 17 Mar. 2023, pp. 37-45, https://doi.org/10.61414/jeti.v5i1.103.

Hur, Jung Won. "Fostering AI literacy: Overcoming concerns and nurturing confidence among preservice teachers." *Information and Learning Sciences* , vol.126, no.1, 2025, pp. 56-74. https://doi.org/10.1108/ILS-11-2023-0170

Karatay, Yasin, and Leyla Karatay. "Automated writing evaluation use in second language classrooms: A research synthesis." *System* (2024): 103332.

Kočková, Petra, et al. "AI Literacy in Teacher Education in the Czech Republic." *Proceedings of The 23rd European Conference on e-Learning*. Academic Conferences International, 2024, pp. 178-86.

Law, Locky. "Application of Generative Artificial Intelligence (GenAI) in Language Teaching and Learning: A Scoping Literature Review." *Computers and Education Open*, vol. 6, 28 Mar. 2024, p. 100174, https://doi.org/10.1016/j.caeo.2024.100174.

Li, Junfei, et al. "Evaluating the Role of ChatGPT in Enhancing EFL Writing Assessments in Classroom Settings: A Preliminary Investigation." *Humanities and Social Sciences Communications*, vol. 11, 27 Sept. 2024, p. 1268, https://doi.org/10.1057/s41599-024-03755-2.

Lin, Shiming, and Peter Crosthwaite. "The Grass Is Not Always Greener: Teacher vs. GPT-Assisted Written Corrective Feedback." *System*, vol. 127, 29 Oct. 2024, p. 103529, https://doi.org/10.1016/j.system.2024.103529.

Long, Duri, and Brian Magerko. "What Is AI Literacy? Competencies and Design Considerations." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 23 Apr. 2020, pp. 1-16, https://doi.org/10.1145/3313831.3376727.

Lu, Qi, et al. "Can ChatGPT Effectively Complement Teacher Assessment of Undergraduate Students' Academic Writing?" *Assessment & Evaluation in Higher Education*, vol. 49, no. 5, 12 Jan. 2024, pp. 616–33, https://doi.org/10.1080/02602938.2024.2301722. Accessed 27 Feb. 2024.

Mishra, Punya, and Matthew J. Koehler. "Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge."

Teachers College Record, vol. 108, no. 6, June 2006, pp. 1017-54, https://doi.org/10.1111/j.1467-9620.2006.00684.x

Mizumoto, Atsushi, and Masaki Eguchi. "Exploring the Potential of Using an AI Language Model for Automated Essay Scoring." *Research Methods in Applied Linguistics*, vol. 2, no.2, 19 Apr. 2023, p.100050, https://doi.org/10.1016/j.rmal.2023.100050.

Ng, Davy Tsz Kit, et al. "Conceptualizing AI Literacy: An Exploratory Review." *Computers and Education: Artificial Intelligence*, vol. 2, 22 Nov. 2021, p. 100041, https://doi.org/10.1016/j.caeai.2021.100041

Page, Matthew J., et al. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews." *Systematic Reviews*, vol. 10, 29 Mar. 2021, p. 89, https://doi.org/10.1186/s13643-021-01626-4.

Parker, Jessica L., et al. "ChatGPT for Automated Writing Evaluation in Scholarly Writing Instruction." *Journal of Nursing Education*, vol. 62, no. 12, 01 Dec. 2023, pp. 721–27, https://doi.org/10.3928/01484834-20231006-02.

Patton, Michael Quinn. *Qualitative evaluation and research methods*. SAGE Publications, Inc, 1990.

Purpura, James Enos. "Learning-Oriented Language Assessment." *The Concise Companion to Language Assessment,* edited by Antony John Kunnan,1st ed., John Wiley & Sons, 2024, pp. 22-41.

Ranalli, Jim, et al., "Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation." *Educational Psychology*, vol. 37, no. 1, 01 Feb 2016, pp. 8-25, https://doi.org/10.1080/01443410.2015.1136407.

Reynolds, Laria, and Kyle McDonell. "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm." *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 8 May 2021, https://doi.org/10.1145/3411763.3451760. Accessed 8 Nov. 2022.

Shabara, Ramy, et al. "TEACHERS OR CHATGPT: THE ISSUE OF ACCURACY AND CONSISTENCY IN L2 ASSESSMENT." *Teaching*

English with Technology, vol. 24, no. 2, 2024, pp. 71-92, https://doi.org/10.56297/vaca6841/lrdx3699/xsez5215.

Shermis, Mark D., and Jill Burstein, editors. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge, 2013, https://doi.org/10.4324/9780203122761.

Shin, Dongkwang, and Jang Ho Lee. "Exploratory Study on the Potential of ChatGPT as a Rater of Second Language Writing." *Education and Information Technologies*, vol. 29, 13 June 2024, pp. 24735–57, https://doi.org/10.1007/s10639-024-12817-6. Accessed 5 Oct. 2024.

Steiss, Jacob, et al. "Comparing the Quality of Human and ChatGPT Feedback of Students' Writing." *Learning and Instruction*, vol. 91, 11 Mar. 2024, p. 101894, https://doi.org/10.1016/j.learninstruc.2024.101894.

Stevenson, Marie. "A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction." *Computers and Composition*, vol. 42, 19 July 2016, pp. 1-16, https://doi.org/10.1016/j.compcom.2016.05.001.

Stryker, Cole, and Mark Scapicchio. "What is Generative AI?" *IBM*, 22 Mar. 2024, www.ibm.com/think/topics/generative-ai. Accessed 18 Feb. 2025.

Su, Yanfang, et al. "Collaborating with ChatGPT in Argumentative Writing Classrooms." *Assessing Writing*, vol. 57, 2 June 2023, p. 100752, https://doi.org/10.1016/j.asw.2023.100752.

Sun, Yachao, and Ge Lan. "A Bibliometric Analysis on L2 Writing in the First 20 Years of the 21st Century: Research Impacts and Research Trends." *Journal of Second Language Writing*, vol. 59, 5 Jan. 2023, p. 100963, https://doi.org/10.1016/j.jslw.2023.100963.

Tate, Tamara P., et al. "Can AI Provide Useful Holistic Essay Scoring?" *Computers and Education: Artificial Intelligence*, vol. 7, 13 June 2024, p. 100255, https://doi.org/10.1016/j.caeai.2024.100255.

Teng, Mark Feng. "'ChatGPT Is the Companion, Not Enemies': EFL Learners' Perceptions and Experiences in Using ChatGPT for Feedback in Writing." *Computers and Education Artificial Intelligence*, vol. 7, 26 July 2024, p. 100270, https://doi.org/10.1016/j.caeai.2024.100270. Accessed 1 Sept. 2024.

Turner, Carolyn E., and James E. Purpura. "Learning-Oriented Assessment in Second and Foreign Language Classrooms." *Handbook of Second Language Assessment*, edited by Dina Tsagari and Jayanti Banerjee, Boston, De Gruyter Mouton, 2016, pp. 255-274. https://doi.org/10.1515/9781614513827-018

Wang, Li, et al. "ChatGPT's Capabilities in Providing Feedback on Undergraduate Students' Argumentation: A Case Study." *Thinking Skills and Creativity*, vol. 51, 6 Dec. 2023, p. 101440, https://doi.org/10.1016/j.tsc.2023.101440

Yamashita, Taichi. "An Application of Many-Facet Rasch Measurement to Evaluate Automated Essay Scoring: A Case of ChatGPT-4.0." *Research Methods in Applied Linguistics*, vol. 3, no. 3, 27 June 2024, p. 100133, https://doi.org/10.1016/j.rmal.2024.100133.

Yancey, Kevin P., et al. "Rating Short L2 Essays on the CEFR Scale with GPT-4." *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023),* July 2023, pp. 576–584, https://doi.org/10.18653/v1/2023.bea-1.49.

Address for correspondence

Jie YANG
Run Run Shaw Creative Media Centre
Department of English
City University of Hong Kong
8 Tat Hong Avenue,
Kowloon Tong
Kowloon
Hong Kong SAR

jyang399-c@my.cityu.edu.hk

Submitted Date: January 08, 2025
Accepted Date: June 19, 2025