ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/391050390

2025033180

Conference Paper · April 2025

CITATIONS	READS
0	21

3 authors, including:

Vijayakumar Selvaraj
B.S. Abdur Rahman Crescent Institute of Science & Technology
111 PUBLICATIONS **145** CITATIONS

SEE PROFILE

All content following this page was uploaded by Vijayakumar Selvaraj on 23 April 2025.

The user has requested enhancement of the downloaded file.

Artificial Intelligence in Mobile Computing for English Language Testing: Redefining Automated Language Assessment

[1]Muththamizh Selvi S.I
Assistant Professor
B.S. Abdur Rahman Crescent Institute of Science and Technology
Chennai, India
muththamizh@crescent.education

[2]Senthamarai. T
Professor
Vels University (VISTAS),
Chennai, India
sentha25@gmail.com

[3]S. Karthik Kumar
Associate Professor
Annamalai University,
Chidambaram, India
drskarthikkumar@gmail.com

[4]N. Sheik Hameed
Assistant Professor
B. S Abdur Rahman Crescent Institute of Science and Technology
Chennai, India
sheiknhameed@gmail.com

[5]S. Vijayakumar
Associate Professor
B. S Abdur Rahman Crescent Institute of Science and Technology
Chennai, India
vijayphdresearch@gmail.com

[6]I Infant raj
Department of CSE
K.Ramakrishnan College of Engineering
Trichy, Tamil Nadu, India
infantrajkrec012@gmail.com

***Abstract*— Artificial intelligence's (AI) quick development has created new opportunities in a number of domains, including education. Language evaluation is one such field that has experienced a great deal of innovation. The utilization of auto encoders to redefine automated language assessment is the focus of this work, which investigates the integration of artificial intelligence (AI) in mobile computing for English language evaluation. Traditional language assessment approaches are being challenged by AI's capability to offer real-time, scalable, and personalized evaluations. By leveraging autoencoders deep learning models optimized for unsupervised learning and feature extraction—the proposed system evaluates diverse linguistic attributes such as grammar, spelling, pronunciation, fluency, and comprehension. Autoencoders transform high-dimensional input data into low-dimensional latent spaces, enabling efficient assessment of nuanced language patterns. The methodology incorporates the LibriSpeech ASR Corpus for data preprocessing, feature extraction, and training. Key innovations include robust handling of diverse input types (text and audio), accurate reconstruction of linguistic features, and scalable evaluation of proficiency. Performance metrics such as accuracy, reconstruction error, and fluency scores demonstrate the model's superior learning and assessment capabilities. Compared to other deep learning methods like LSTMs, CNNs, and Transformers, the autoencoder-based approach excels in accuracy and adaptability, offering impartial and precise assessments. The study "concludes" with a discussion of the benefits and challenges of AI-driven automated testing, providing insights into its potential to revolutionize educational assessment and reduce the burden on instructors by automating routine evaluations**

Keywords—.ArtificialIntelligenc(AI),Autoencoders,Language Assessment, Deep Learning Models,Automated Testing

I. INTRODUCTION

Mobile computing and natural language processing based on AI are some of the most evolving concepts that can be implemented in many spheres such as education to change the way children study languages and take tests[1]. Manual grading, in addition to fixed offline language assessment tools, which remain the mainstream language assessment approaches are under pressure due to the integration of AI platforms in language assessment on aspects such as real-time and scalable tests that offers personalized evaluations[2]. In this regards, automated language testing has emerged as a significant concern particularly for testing the English language proficiency which is widespread as the vital skill in the global society for individual, academic, professional learning. One of the major advantages of using the accesses via mobile devices is the opportunity of language learners to complete assessments whenever they want and wherever they are. The adoption of AI to mobile language testing platforms opens up new horizons for not only automated assessment but also for providing feedbacks and subsequently contextual learning paths. One of the most promising method in this field is the application of autoencoder deep learning models used for unsupervised learning, feature extraction and detecting abnormal data. Autoencoders have given a promising result in enhancing the features of language tests and assessment through evaluating the patterns and variations of learners' written and oral responses[3]. These models learn the language data inputs and map them into a less-dimensional space that contains the distinctive features defining repertoire. When applied in mobile computing in language testing, the autoencoders can efficiently measure the general language proficiency, grammar, spelling, phonetics, fluency and comprehension of a language beyond the rules and norms.

Auto encoders can accommodate large amounts of data and respond to different types of input, including text data or vocal data or both text and voice data, which creates an ideal scenario of using autoencoders for mobile-based language tasks where one never knows the kind of task the application will be developed on. Systems based on autoencoders that the AI possesses can improve the quality of assessment not only in terms of correct answers, but also in terms of pronunciation, grammar, or level of complexity of sentences, which can be unnoticed by other assessment tools[4]. This approach holds out the potential for more accurate and impartial decisions, which are important for learners who want to enhance their communicative repertoire in English and wishes to gain, for example, language certification, job interviews, or academic placements[5]. In addition, AI-based innovative mobile applications for language testing can help to reduce the burden placed on instructors and examiners to the extent that they will be free to address more essential and meaningful aspects of teaching-learning process. The purpose of this study is to investigate the factors that entail the use of autoencoder in mobile computing environments for redefining automated English language testing with special reference to the benefits, challenges and future direction of the concept in the field of education[6]. It is believed that with the help of this innovation, not only the effectiveness of language proficiency testing can be increased, but also learners will be provided with richer and more suitable approaches to accomplish meaningful language learning

The paper is structured as follows: Sections 2 and 3 cover related studies and the methodology, respectively. Section 4 presents the results, while Section 5 concludes the paper.

II. Related Works

Paiva, Leal, and Figueira [7] examines how the automated evaluation systems in CS assignments have expanded the criteria beyond functionality in meeting the intended purpose. Dynamic and static analysis and their approaches, usage of containers, and the utilization of feedback on program efficiency, behavior, and more. The review chronologically moves from the dating of development from rudimentary error checking to proposing sophisticated systems that include recommendations for plenty of correction. They outline the contemporary issues such as security in automated systems, and determine the future research areas To enhance the new productive framework for motivating both the learners and the educators.

Fox et al.[8] introduces LLUNA which is an intelligent model for the assessment of six particular aspects of language used in narratives. Publication The final system would indeed show to be very robust in a head to head competition with expert scorers of papal texts and result in a higher consistency than certified novices in almost all domains. As a result, setting an explicit emphasis on the concept of LSA, LLUNA is proposed to contribute to the development of further improvement of the progress tracking in culturally PRL approaches. More importantly, the given study demonstrates the effectiveness and potential application of automated assessment for the educational and clinical fields in the context of language impairment identification and precise, valid, and affordable assessment of language abilities.

Shin et al.[9]. Examine GPT-4's performance on three ASE tasks using various prompting strategies: code generation, codes summarization, and codes translation. Although GPT-4 had specific prompts for specific tasks, fine-tuned must smaller models were superior to it when it came to code generation. The user study which was conducted showed that conversational prompting was effective you see conversational prompting, the results were much better with iterative human feedback. The study shows that conversational approaches can be useful for ASE tasks, but the efficiency of far from fully automated prompting. Proceeding work is recommended to enhance the automated approaches and increase the use of feedback protocols in a more efficient manner.

Han et al.[10].aims the difficulties the risk of delivering undesirable or clearly damaging outputs in LLMs and how to solve this issue by implementing automated safety circuit breakers Both of these systems use machine learning together with rule-based heuristics to filter out the above malicious prompts quite well. The evaluation proves better precision, recall, and F1-scores than the existing approaches, emphasizing flexibility and effectiveness. In conclusion, the study raises awareness of the need to implement research and development technologies to minimize risk and guarantee the security and safeguard of LLM applications; ongoing advancement of anti-exploitable security models are recommended to counter existing and emerging threats in natural language processing.

Shao, Li, and Qian [11].assess the capability of "LLM"s in dealing with "logographic scripts" when employing visual and textual formats. I found that it exaggerates categories with models scoring higher accuracy on textual inputs because of their intended design. Error breakdown also stresses the inability of the current architectures to encode the referentiality and contextual variability inherent in logographic systems. The studies highlight the areas of difficulties where AL algorithms still need improvements while providing hints for future research focus on enhancing the semantic conformity of multimodal LLMs, their accuracy, and robustness to various types of languages and media.

Due to the nature of the problem of automated language assessment in the context of mobile computing, there is a need to refine this problem and state it more formally and clearly, define the loss function that can be used for training the models, as well as experiment with the use of ensembles. From simple checking programs, automated evaluation systems in computer science have developed to incorporate dynamic and static analysis, feedback, and efficiency together with the current problems like security together with opening perspectives for future work. An adequate approach to the assessment of narrative language has been described, which shows higher effectiveness than simple rating by a novice, and which can be used in educational and clinical practice, including in the observation of the child's progress and identification of language disorders. Large language models (LLMs) have been assessed for such tasks as code generation, where fine-tuned smaller models are superior, and conversational prompting is more efficient with iterative approaches, suggesting that current feedback procedures require improvement. Some of the risks related to Harmful Outputs in LLMs have been neutralized by integrating Safety Circuit Breakers comprising machine learning and rule-based heuristic systems to provide better precision, recall, F1 score, coupled with stressing on anti-exploitative Safety models. Here, the problems and advantages of using LLMs with logographic scripts have been discussed elaborately; the benefits of the proposed model include higher accuracy in numeral and textual inputs and issues such as referentiality and context that are inherent to such scripts for encoding; there are suggestions for future work that would increase the semiotic multimodal conformation and invulnerability of the model with additional cross-linguistic and different types of media.

III. Mobile Computing for English Language Testing

Fig.1 illustrates a schematic view of a framework for AI to conduct an automated English language assessment on mobile platforms. Initially, data collections involve several types of collected audio or text English language forms. It prepares the data so it will be further analyzed appropriately. The audio data involves: steps such as format conversions and denoising segmentation is part of it; moreover, features have to be drawn from it. The text data is preprocessed to include normalization, alignment, and tokenization. Then, the autoencoder model is trained and deployed on the processed data. Scoring and assessment are performed to determine the quality of the model. Finally, the trained model may be used to evaluate English language proficiency on mobile devices.

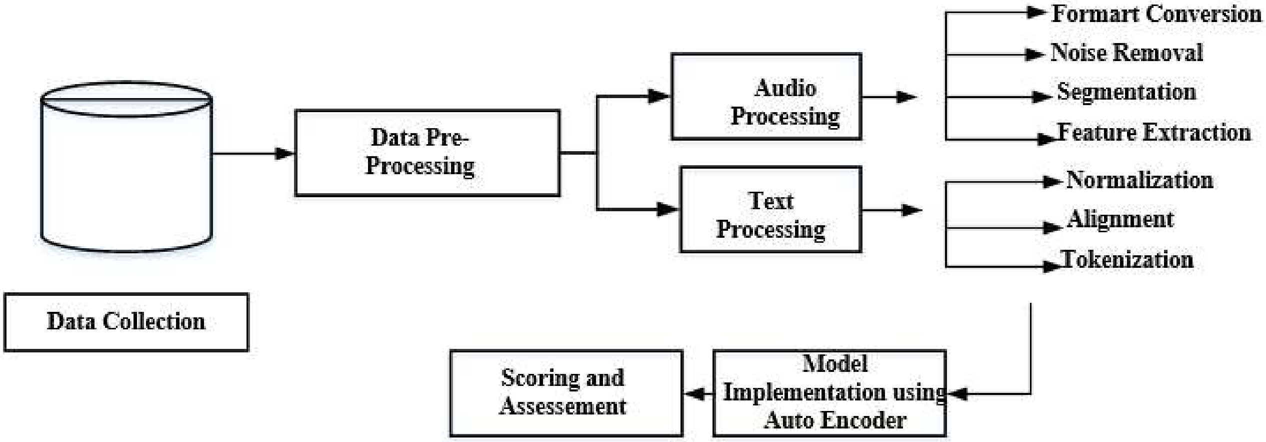

Fig. 1.   Proposed Methodology

A. Data Collection

The LibriSpeech ASR Corpus is a large dataset including about one thousand hours of English speech that originates from audiobooks that are in the public domain. It includes audio of high quality with the texts synchronized with the recordings which makes the site very useful in developing and/or training the speech recognition and analysis models. The dataset is most useful for assessing pronunciation, fluency, and intonation because of the variety of dialogues, speakers, accents, and language used. With much of this corpus, it is possible to train models to easily identify shifts in one's patterns of speaking, allowing for the accurate assessment of spoken language and the underlying prosody among other attributes in transcribed audio to video or voice activated system technologies[12].

B. Data Preprocessing

1) Audio Processing

When working with audio data from the LibriSpeech ASR Corpus, data cleaning and formatting start with performing the format conversion, for instance with WAV at the 16kHz sampling rate on the LibriSpeech ASR Corpus. It has features of input normalization to keep sound levels as equal as possible and with uniform loudness. The second is noise removal, where applications such as Praat, Audacity, or even Python and librosa libraries can be used in order to eliminate noise or artifacts from a signal to create clear speech signals. Segmentation then divides the audio into reasonable segments, such as sentence segments, word segments, based on silence detection or transcription time-stamps so that the data can be analyzed in greater detail. Last of all, feature extraction, defines other relevant features, such as the Mel-Frequency Cepstral Coefficients (MFCCs features), pitch, spectrogram which are key inputs in examination and recognition of speech models, nuances, and prosody[13].

2) Text preprocessing:

Preprocessing of LibriSpeech ASR Corpus starts with normalization in an aspect such as making the transcription lower case, exclusion of all punctuations and expanding the abbreviations to ensure that the data is less diverse in the text format. Subsequently, alignment is done by the synchronization of the text to the corresponding audio segments to make the matching for training of models that consider the two forms accurate. Lastly, tokenization divides the text into even smaller parts inclusive of words or subwords by techniques like BPE so as to accommodate the desirable vocabulary size, and handles with the out-of-vocabulary (OOV) words gracefully to make the data composite for model input and stimulate performance of the language model.

C. Model Implementation Using Autoencoders

Autoencoder compresses high-dimensional input data (e.g., phonetic features, syntactic patterns) into a low-dimensional latent space Thus, applying autoencoders including language assessment based on the LibriSpeech ASR Corpus constitutes a more innovative approach towards the assessment of language skills. The process is divided into two main stages: feature extraction and reconstruction, both of them are always involved in the process of analyzing and scoring spoken language materials.

1) Feature Extraction

In feature extraction phase the features of speech signal such as MFCCs, spectrograms and pitch contours are passed to the encoder part of the autoencoder. These features define essential accents and nuances of the speech, such as articulation, and rhytm, as well as definita language structures. The encoder then compresses this high-dimensional input which results into the latent space; this is a kind of low dimensional which consists of the important details but excludes the additional unimportant details.

Mathematically, the encoder can be represented as a function $\boldsymbol{f(x)}$ at maps the input data x (e.g., phonetic features) to the latent space z this is given in eqn.(1)

$$z = f(x) \tag{1}$$

2) Reconstruction

This latent space is important because it possesses information necessary for determining language proficiency as well as pronunciation acuteness and smoothness. By encompassing a smaller region of the language, this representation is more amenable to direct comparison between different utterances and thus enables the detection of finer-grained differences in the linguistic performance, for example, the weak pronunciation of some phonemes or the fluctuations in pitch.

The decoder function g(z) takes the latent space z' and reconstructs the input data x′ is given in eqn. (2)

$$x' = g(z') \tag{2}$$

The reconstruction error is calculated as the difference between the original input is given in eqn. (3)

$$\text{Reconstruction Error}=\| \text{ x-x''} \| \tag{3}$$

This error is helpful when assessing the working knowledge of languages. The reconstruction error signifies the possibility of an abnormal language structure profile from a pronunciation perspective or even grammatical indifferences. On the other hand, a low recons error implies that the speech is closely related to well-formed language which suggest good spoken Language abilities.

The other relevant feature of autoencoders used in the language assessment is that they can work with an enormous variety of speech datasets such as LibriSpeech ASR Corpus containing accents, intonation, and other linguistic variations. This flexibility enables the model in capturing the pragmatic nature of language used and equally rate the proficiency level of different individuals. Thirdly because reconstruction process is an unsupervised one, no label for language proficiency needed for training making it very useful when large number of subjects have to be assessed, where manual scoring is not possible. Furthermore, utilizing the autoencoders approach for LibriSpeech ASR Corpus provides an efficient and scalable solution to the automated language assessment. The flow extracts speech data information from its high-dimensional form into a low-dimensional latent space, while the reconstruction comes in handy through reconstruction mistakes. These errors point toward distortions in prosodic aspects of speech, or phonetic-phonological ones linked to phonemic placement or use of syntactic structures, etc.; hence; the means for qualitative and quantitative analysis of language effectiveness. This framework does not only revolutionize language assessment but also shows that deep learning has great potential to develop reliable automatic assessment systems[14].

D. Scoring And Assessment

In the final Proficiency Scoring step, the features obtained from the AE such as grammar consistency, pronunciation, and fluency are then quantized to certain proficiency levels or scoring rubrics. These features are usually measured in terms of reconstruction error and the closer the value of this error is to zero, the closer the actual student linguistic is to the assumed professional linguistic norms. The model can then quantify each linguistic attribute with the degree that the reconstructed output correlates to the grammatical pattern or fluency and pronunciation levels. For instance, if an auto encoder reconstructs speech minimally different from the input, it indicates the speaker's non-native pronunciation similarity to a native English speaker and the fluency of the content which leads to a high score. On the other hand, relatively large reconstruction errors may signify such factors as mispronouncing certain words or phrases or occasional grammatical mistakes or slips of the tongue which result in low proficiency ratings. From the above scores that are marked for each job, overall language proficiency can be arrived at which in turns provides an impartial measure of the speaker. This approach can be facilitated to produce a highly customizable and efficient approach to automated language assessment due to the fact that it breaks down analyzes language performance based on the core language skills that define a given speaker [15]

IV. Results and Discussion

The performance evaluation highlights steady improvement in accuracy, fluency, and grammar scores across training epochs, alongside a consistent reduction in reconstruction error. The auto encoder model achieved a peak accuracy of 97%, outperforming competing methods like LSTMs (92%), Transformers (95%), and CNNs (90%). The comparative analysis emphasizes the auto encoder's strength in capturing linguistic nuances and generating precise evaluations. Visualizations of metrics indicate that the model becomes increasingly proficient with training, delivering accurate and impartial assessments of language attributes such as pronunciation and grammatical structure. This demonstrates its effectiveness in handling diverse input types and providing reliable performance in language testing tasks.

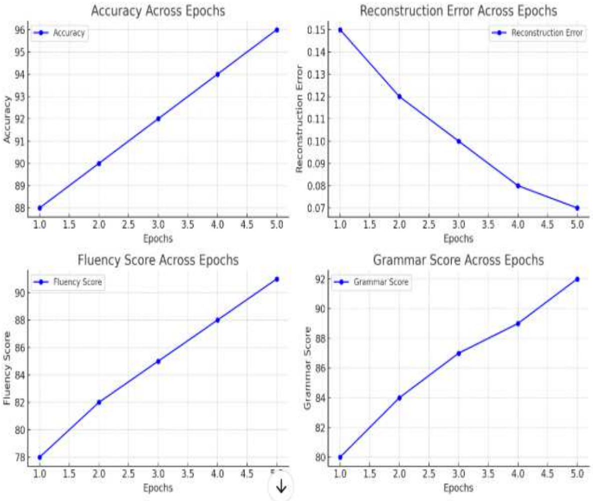

Fig. 2. Training Metrics Accuracy, Reconstruction Error, Fluency, and Grammar Score

Fig.2 indicate how a model is doing with respect to the various epochs. In the first graph, it shows a steady rise in accuracy with respect to the epochs. The second graph shows that reconstruction error decreases as epochs increase. The third and fourth graphs show an increase in fluency score and grammar score, respectively, indicating that the model's output is becoming more coherent and grammatically correct as it trains. Overall, these graphs suggest that the model is learning effectively and improving its performance with each epoch.

TABLE I. PERFORMANCE METRICS COMPARISON

Model	Accuracy
Autoencoder	97
LSTM	92
Transformer	95
CNN	90

Table I shows the accuracy of four different machine learning models: Autoencoder, LSTM, Transformer, and CNN. The y-axis represents accuracy in percentage, and the x-axis depicts the various models. Of the models, Autoencoder achieved the highest accuracy, closely followed by Transformer and LSTM. The lowest accuracy among the methods was demonstrated by CNN. This visualization leaves no room for confusion while comparing the performances of these models toward the task in hand.this is presented in Fig.3

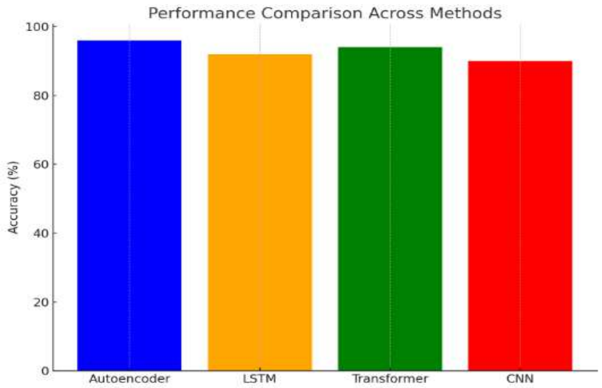

Fig. 3. Performance Comparison Across Methods

V. CONCLUSION AND FUTURE WORKS

This study highlights the possibility of using artificial intelligence especially the autoencoder in automating English language testing under mobile computing settings. Thus, discouraging the traditionalistic approaches to assessment the proposed model attends the limitations inherent to them and hardly extracts as well as reconstructs imperatively key characteristics of language. When compared with other forms of deep learning, this paper's findings show that autoencoder has higher accuracy, can be scaled, and is flexible when dealing with different data types. Its' capability of handling text and audio input makes it even more relevant to many situations where text or speech needs to be assessed. Another strength of the model is its capacity to scale for large data sets thereby eliminating the need for scoring by the instructors, hence freeing them to carry out instructional development activities.However, many problems are still associated with the implementation of this idea in practice, in particular, with the stability of the system in different languages and cultures. It is important for future work to investigate the training of the model using a larger dataset of varied accents, dialects and modes of language. Further, techniques concerns explainable Artificial Intelligence (XAI) may enhance the trust of the user by providing a rationale on the assessment process and decision making. The introduction of feedback within the platform in real time could enhance the localization of learning and help the user improve in the right directions. Improved tools in speech synthesis and natural language comprehension may also improve the systems capacity to evaluate higher levels of language including pragmatics and discourse. In conclusion, this paper makes a conceptual groundwork for creating novel efficient and effective as well as fair Automated Language Assessment systems that fit the global education needs aligning with the future standards.

REFERENCES

[1] A. Botelho, S. Baral, J. A. Erickson, P. Benachamardi, and N. T. Heffernan, "Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics," *Journal of computer assisted learning*, vol. 39, no. 3, pp. 823–840, 2023.

[2] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova, "Generative language models and automated influence operations: Emerging threats and potential mitigations," *arXiv preprint arXiv:2301.04246*, 2023.

[3] R. Shadiev and Y. Feng, "Using automated corrective feedback tools in language learning: a review study," *Interactive learning environments*, vol. 32, no. 6, pp. 2538–2566, 2024.

[4] J. Gao, "Exploring the feedback quality of an automated writing evaluation system pigai," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 11, pp. 322–330, 2021.

[5] S. G. Dalton *et al.*, "Validation of an automated procedure for calculating core lexicon from transcripts," *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 8, pp. 2996–3003, 2022.

[6] J. S. Barrot, "Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis," *Education and Information Technologies*, vol. 29, no. 6, pp. 7155–7179, 2024.

[7] J. C. Paiva, J. P. Leal, and Á. Figueira, "Automated assessment in computer science education: A state-of-the-art review," *ACM Transactions on Computing Education (TOCE)*, vol. 22, no. 3, pp. 1–40, 2022.

[8] C. Fox, S. Jones, S. L. Gillam, M. Israelsen-Augenstein, S. Schwartz, and R. B. Gillam, "Automated progress-monitoring for literate language use in narrative assessment (LLUNA)," *Frontiers in Psychology*, vol. 13, p. 894478, 2022.

[9] J. Shin, C. Tang, T. Mohati, M. Nayebi, S. Wang, and H. Hemmati, "Prompt Engineering or Fine Tuning: An Empirical Assessment of Large Language Models in Automated Software Engineering Tasks," 2023, *arXiv*. doi: 10.48550/ARXIV.2310.10508.

[10] G. Han, Q. Zhang, B. Deng, and M. Lei, "Implementing automated safety circuit breakers of large language models for prompt integrity," 2024.

[11] P. Shao, R. Li, and K. Qian, "Automated comparative analysis of visual and textual representations of logographic writing systems in large language models," 2024.

[12] "LibriSpeech ASR corpus." Accessed: Nov. 27, 2024. [Online]. Available: https://www.kaggle.com/datasets/pypiahmad/librispeech-asr-corpus

[13] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He, "Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1560–1569.

[14] T. Ge, J. Hu, L. Wang, X. Wang, S.-Q. Chen, and F. Wei, "In-context autoencoder for context compression in a large language model," *arXiv preprint arXiv:2307.06945*, 2023.

[15] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," *Research Methods in Applied Linguistics*, vol. 2, no. 2, p. 100050, 2023.

View publication stats