A Study on Automatic Scoring and Diagnostic Feedback Generation of Vocational College English Application Writing Based on a Large Language Model

Lanna He[1]
[1]Public Education Department, Linyi Vocational University of Science and Technology, Linyi 276000, Shandong, China
Corresponding author: Lanna He (e-mail: helanna19830418@163.com)

ABSTRACT This paper examines the use of large language models to construct an automatic scoring and diagnostic feedback generation system for vocational college English application writing. First, a corpus of vocational college English application writing texts covering multiple genres is constructed. Based on manual scoring and multidimensional feature annotation, a scoring rubric aligned with vocational college English teaching practice is developed. Second, a finely tuned pre-trained language model is used to implement an automatic scoring module and a diagnostic feedback generation module. The scoring module adopts a BERT-based architecture for discriminative scoring, while the feedback module uses a generative model to produce structured diagnostic comments. Experimental results show that the mean weighted consistency coefficient between automatic and manual scoring reaches 0.87. Diagnostic feedback is positively evaluated in terms of content accuracy, personalization, and teaching usefulness. The system can generate overall evaluations, dimensional diagnoses, and targeted improvement suggestions for applied writing. The study provides an effective intelligent auxiliary tool for English writing instruction in vocational colleges and supports scalable formative assessment.

INDEX TERMS large language model, vocational english, application writing, automatic scoring, diagnostic feedback

I. INTRODUCTION

STUDIES on the crossroads of vocational education and language teaching are registering a trend of diversifying and deep development. The interest of vocational college English teaching is to foster the practical application of language use skills in students, and as a functional manifestation of the language output competence, applied writing takes a central place in teaching assessment.

Nonetheless, a vocational college can be characterized by certain practical issues, including high class enrolment, unequal English language proficiency, and excessive teacher assessment, which can cause an extensive period of writing feedback and a low level of individualization, which does not make it easy to address the needs of students to provide timely learning diagnostics and feedback on how to improve it.

In the recent past, massively scaled language models have proven to have strong performance on natural language processing tasks, like text generation, semantic comprehension, and text critique, creating new technical directions toward scoring automatic writing and generating feedback. Existing studies (including domestic and international research) have used large language models to English writing assessment, although at this point they largely center around academic or general writing of English, and the research on automatic scoring as well as the generation of diagnostic feedback on applied writing styles in vocational colleges is relatively poor.

The vocational applied writing has well defined stylistic conventions, communicative objectives as well as structure requirement and scoring dimensions and comments vary largely compared to general writing. Hence, the development of an automatic scoring and diagnostic feedback system of vocational college English writing is not merely a technological necessity, but as well a practical need in the teaching process.

The paper focuses on the specific context of English writing in vocational colleges, developing an automatic scoring and diagnostic feedback generation system using large language models, and validates its efficiency and pedagogical relevance through controlled experiments in teaching via experiments, which attempts to offer an intelligent resource

to facilitate teaching English writing in vocational colleges.

II. RELATED WORK

Research on the intersection of vocational education and language teaching is becoming increasingly abundant, providing diverse perspectives for the reform of English teaching in higher vocational colleges. Ismailia [1] explored the applicability of project-based learning in vocational English teaching. Idham et al. [2] reviewed the opportunities and challenges brought about by the application of AI in language education.

Husnaini [3] Develop self-esteem-oriented microteaching materials, using research and development (R&D) methods, through stages such as needs analysis, material design, expert verification, trial teaching and effect evaluation. Sawalmeh and Dey [4] explored the rapid growth in demand for English oral teachers in the context of globalization. Ghafar [5] concluded that ESP teaching has unique value in meeting the needs of specific disciplines or professions, but there are still significant difficulties in the accuracy of needs analysis, the adaptability of teaching materials and resources, and teacher professional development.

Based on Vygotsky's "zone of proximal development" theory and scaffolded teaching strategy, Alghamdy [6] examined the actual application of English teachers' classroom discourse practice. Lee et al. [7] used experimental design methods to construct an automatic question generation system based on prompting engineering and compared it with traditional manual question generation and existing automatic question generation systems. Uyun [8] implemented a variety of oral teaching strategies in English classrooms and evaluated the effectiveness of the strategies through classroom observation, student feedback and oral tests.

Çelik and Memduhoglu [9] took the Turkish English teacher education project as the research object and systematically evaluated its curriculum design, implementation effect and training quality. Muhajir and Symsu [10] explored the creative application of Pecha Kucha, a rapid presentation format, in English teaching, aiming to improve students' expression ability and classroom participation.

This study focuses on the specific scenario of vocational college English application writing, and constructs an automatic scoring and diagnostic feedback generation system based on a large language model, aiming to explore the technical feasibility and teaching applicability, and provide intelligent auxiliary means for vocational college English writing teaching.

III. METHODS

A. DATA COLLECTION AND CORPUS CONSTRUCTION

This study collected 1820 application writing texts from non-English major students at a vocational college during the 2023-2024 academic year. The texts covered six common application writing styles: letters, notices, notes, application letters, complaint letters, and emails. The distribution of texts for each style was relatively even: 320 letters, 305 notices, 295 notes, 310 application letters, 300 complaint letters, and 290 emails.

Each essay was independently graded by two teachers with over five years of experience teaching English in vocational colleges. The grading was based on a 100-point scale, and the average of the two teachers' scores was used as the final human evaluation label. During the grading process, teachers followed standardized grading criteria. The grading dimensions included content completeness, language accuracy, stylistic conformity, and structural clarity, with each dimension having a weight of 0.35 for content completeness, 0.30 for language accuracy, 0.20 for stylistic conformity, and 0.15 for structural clarity.

To ensure consistency in scoring, teachers were trained on scoring criteria before the formal scoring process, and 10% of the samples were selected for consistency testing during the scoring process. The results showed that the intragroup correlation coefficient between the scores of two teachers was 0.92, indicating that the scoring had high consistency.

In addition to manual scoring, the study also performed multi-dimensional feature annotation on all texts, including text length, lexical diversity, syntactic complexity, error type distribution, and stylistic feature conformity, providing basic data support for subsequent model training and evaluation [11,12].

The corpus was divided into training set, validation set, and test set in an 8:1:1 ratio. The training set contained 1456 texts, the validation set contained 182 texts, and the test set contained 182 texts. During the partitioning process, it was ensured that the proportion of texts of each genre in the three datasets remained consistent to avoid the impact of uneven genre distribution on model evaluation.

B. CONSTRUCTION OF AUTOMATIC SCORING MODEL

The automatic scoring model is built upon a pre-trained BERT-based architecture, rather than a generative GPT model, to leverage BERT's superior bidirectional semantic encoding capabilities for discriminative tasks. While GPT models excel in text generation, BERT provides more stable and precise numerical output for regression-based scoring tasks by capturing nuanced linguistic features. The diagnostic feedback module, conversely, utilizes the GPT series to exploit its advanced natural language generation strengths, creating a hybrid pipeline that prioritizes precision in assessment and fluency in feedback.

In the domain-adaptive pre-training phase, texts from vocational college English textbooks, student writings, and model application documents were collected to continuously pre-train the BERT model, enabling it to better adapt to the language style and stylistic features of vocational college English application documents. Continuous pre-training employed a masked language model task, with three training epochs, a learning rate of 2×10^{-5}, and a batch size of 32.

In the scoring task fine-tuning phase, human ratings from the training set were used as supervision signals. A fully connected layer was added to the top layer of the pre-

trained model to map the model's hidden representations to the scoring values. During fine-tuning, mean squared error was used as the loss function, AdamW was selected as the optimizer, the learning rate was set to 3×10^{-5}, the training epochs were five, and the batch size was 16. To prevent overfitting, an early stopping mechanism was introduced during training; training was stopped when the validation set loss no longer decreased after two consecutive epochs.

The model input is in the form of "[CLS]text content[SEP]", and the hidden layer output corresponding to the "[CLS]" position is used as the overall text representation. In addition to the basic scoring model, the study also constructed a multi-dimensional scoring model to predict four dimensions: content completeness, language accuracy, stylistic norms, and structural clarity. Based on the basic model structure, the multi-dimensional scoring model extends the top fully connected layer into four independent units.

While these dimensions share the parameters of the underlying encoder to capture universal linguistic features, the model utilizes a multi-task learning (MTL) framework designed to mitigate task interference. We acknowledge that tasks like "stylistic conformity" and "language accuracy" require different feature sets; however, the shared encoder facilitates a holistic semantic understanding, while the independent top layers allow for dimension-specific gradients. This approach balances the synergy between related writing traits while preventing the specific requirements of one dimension from overshadowing another, and the loss function is the sum of the mean squared errors of the four dimensions.

The weights of each dimension are consistent with the weights of the dimensions in the scoring rules. After the model is trained, the parameters are tuned on the validation set, and finally the performance is evaluated on the test set [13,14].

C. CONSTRUCTION OF DIAGNOSTIC FEEDBACK GENERATION MODEL

The diagnostic feedback generation module is based on the GPT series architecture, integrated with the BERT-based scoring module into a unified pipeline. In this framework, the multi-dimensional scores and feature representations generated by the BERT model serve as structured context for the GPT prompt. Specifically, the feedback generation task involves the model producing natural language diagnostics by synthesizing the raw student text with the specific scoring outputs (content, language, style, structure) provided by the preceding scoring module to ensure consistency between the grade and the feedback.

The feedback generation model is constructed in three phases, which include prompt template design, fine-tuning of instructions, and maximization of feedback quality. The study formulated structured input prompts with the qualities of vocational college writing of English applications at the prompt design phase. The prompts will consist of the text of writing by the student, the type of writing genre the student is applying, the points gained in each dimension, and the marks of an error. The immediate template transforms the above into the natural language instructions and directs the model to produce structured responses.

The feedback output framework is created to have three sections overall assessment, dimensional diagnosis and improvement recommendations. The overall evaluation section gives a concise overview of the overall writing work; the dimensional diagnosis section identifies the specific problems and gives examples in four areas: content completeness, language accuracy, stylistic conformity, and structural clarity; the improvement suggestions section gives the specific modification suggestions and examples.

In the process of data collection, two teachers were also requested to write structured feedback to each text. The feedback was structured based on the overall evaluation structure, the dimensional diagnosis structure and improvement recommendations. One thousand eight hundred and twenty teacher feedbacks were gathered with an average of 215 words of length of the feedback.

A monitored fine-tuning approach was used in the fine-tuning process, whereby student texts and genre types were used as input and teacher feedback used as a desired output to train the model so as to produce diagnostic feedback that could match the format requirements.

A stage on the feedback quality optimization was presented whereby a reinforcement learning strategy was implemented to maximise the quality of the produced feedback. A feedback quality grader was made that would automatically rate the produced feedback on three scales namely accuracy in the content, personalization, and pedagogic applicability. The assessor was trained using manually labelled feedback on quality of feedback. During the reinforcement learning phase, the evaluator score was regarded as a reward signal and a proximal policy optimization algorithm was used to further adapt the model and enhance the quality of the feedback produced.

IV. RESULTS AND DISCUSSION

A. PERFORMANCE EVALUATION OF THE AUTOMATIC SCORING MODEL

Figure 1 provides the performance evaluation outcomes of the automatic scoring model on the test set. To ensure clarity, the primary success metric used is the Quadratic Weighted Kappa (QWK), referred to here as the mean weighted consistency coefficient, alongside the Pearson correlation coefficient and root mean square error (RMSE). The QWK (Consistency Coefficient) is used to determine the similarity between model scores and human scores on a scale of 0 to 1. Pearson correlation coefficient is a measure of linear relationship between the score in the model and the human score. The RMSE is used to evaluate the human score of a model on the average deviation.

When comparing the results of the evaluation of various genres, the most consistent were letters, and the average

weighted consistency coefficient was 0.89. Coefficients of 0.87 and 0.86 followed notification and emails, respectively. Consistency was slightly less with memos, application letters and complaint letters, which stood at 0.84, 0.85 and 0.84 respectively.

The variation in consistency in the scoring of the various genres could be linked to the length of the text and the standardization of the structure. There is high level of structural standardization and fixed format of letters, and models are more easily taught their scoring patterns. On the contrary, notes are characterized by higher linguistic flexibility and informal expressions.

The comparative difficulty in rating notes arises not from text length, but from the inherent linguistic ambiguity and the high variability of "correct" responses in informal contexts, which poses a challenge for the model to extract standardized features compared to more structured genres like letters.

The results indicate that the multi-dimensional scoring model is superior to the basic scoring model in all types of texts, and therefore, the multi-dimensional scoring model enables the model to express the text quality aspects in a more accurate manner, thus enhancing the accuracy of the overall score.

In the case of memos and complaint letters, consistency in scoring is not very high. The reason is that memos are brief and they contain a more relaxed language and there is therefore a certain amount of subjectivity in providing linguistic accuracy and style conventions when scoring the memos. Complaint letters, on the other hand, incorporate complex sentiments and varying degrees of formality.

To address this, the model's domain-adaptive pre-training phase included a sentiment-weighted objective, allowing the encoder to recognize emotional intensity as a key indicator of "stylistic conformity." While subjectivity among human raters remains high, this specialized tuning enables the model to align its scores with the pragmatic expectations of complaint-based communication.

B. DIAGNOSTIC FEEDBACK QUALITY ASSESSMENT

The diagnostic feedback quality test was a blend of human and automated testing. The test set was randomly chosen to provide sixty texts. Model-generated and teacher-written feedback was used concurrently in each text. Blind assessment of both types of feedback was done by the five vocational college English teachers. The dimensions on which it was evaluated were content accuracy, personalization and the pedagogical applicability, where a score of 1 indicated very poor, and 5 very good. Table 1 provides the results of the evaluation.

The findings imply that the model-generated feedback scored an average of 4.32 in content accuracy, thus, showing that the model-generated feedback was quite similar to the teacher in terms of accuracy with some mistakes or inappropriate utterances present. At the personalization dimension, the model generated feedback was an average of 4.18 and the teacher feedback was 4.35 on average. The rather minor gap between the two suggests that the model is able to produce personalized feedback related to the particular situation of the student texts, as opposed to delivering stereotyped and template-driven content.

Regarding teaching practicability, the model-generated feedback had an average of 4.25, whereas, the teacher feedback had an average of 4.62 with a significant difference. Teacher feedback showed that it has a more significant benefit in teaching practicality, which was more likely to merge students actual levels with the learning objectives to provide particular and practical improvement advice.

Further examination of the results showed that the model did not correctly formulate the exact issues in the text when the improvement suggestions were provided or that the suggestions were so generalized and non-specific.

The evaluation of feedback was done by an automated method of evaluation in addition to the human evaluation. This automated assessment was founded on a feedback quality evaluator trained on a manually labelled feedback quality dataset, which had the capability of scoring feedback on three dimensions. The results of the automated scoring were compared to the results of human scoring, and the two were estimated and correlated to ascertain the quality of the feedback evaluator.

The findings revealed that Pearson correlation coefficients existed between the scores given by the feedback quality evaluator and human scores on the three dimensions of content accuracy, personalization, and pedagogical applicability, which were 0.81, 0.78, and 0.79, respectively, which indicated that the evaluator had the ability to reflect feedback quality.

On this basis, the evaluation of the quality of feedback received was done in the feedback quality evaluator to perform batch assessment of the feedback obtained on the test set as a whole. The results of the evaluation revealed that the model generated feedbacks of the 182 test texts in the test set had average scores of 4.28, 4.15 and 4.21 on the three dimensions of content accuracy, personalization and pedagogical applicability respectively which were more or less comparable with the human evaluation results.

As far as the feedback in various genres is concerned, the letters and notices got scored more often, and the notes and complaint letters got scored much lower. This pattern of distribution is also in line with the performance of the automatic scoring model.

C. ERROR ANALYSIS AND FEEDBACK TO VERIFY IMPROVEMENT EFFECTS

The teaching efficacy of the diagnostic feedback was also checked by undertaking a study where the effect of improvement was found to be valid. A corpus was used to select 120 medium-quality texts (having a score of 65 to 75 points) in an experimental and control group (with 60 texts each). The experimental group students were provided with diagnostic

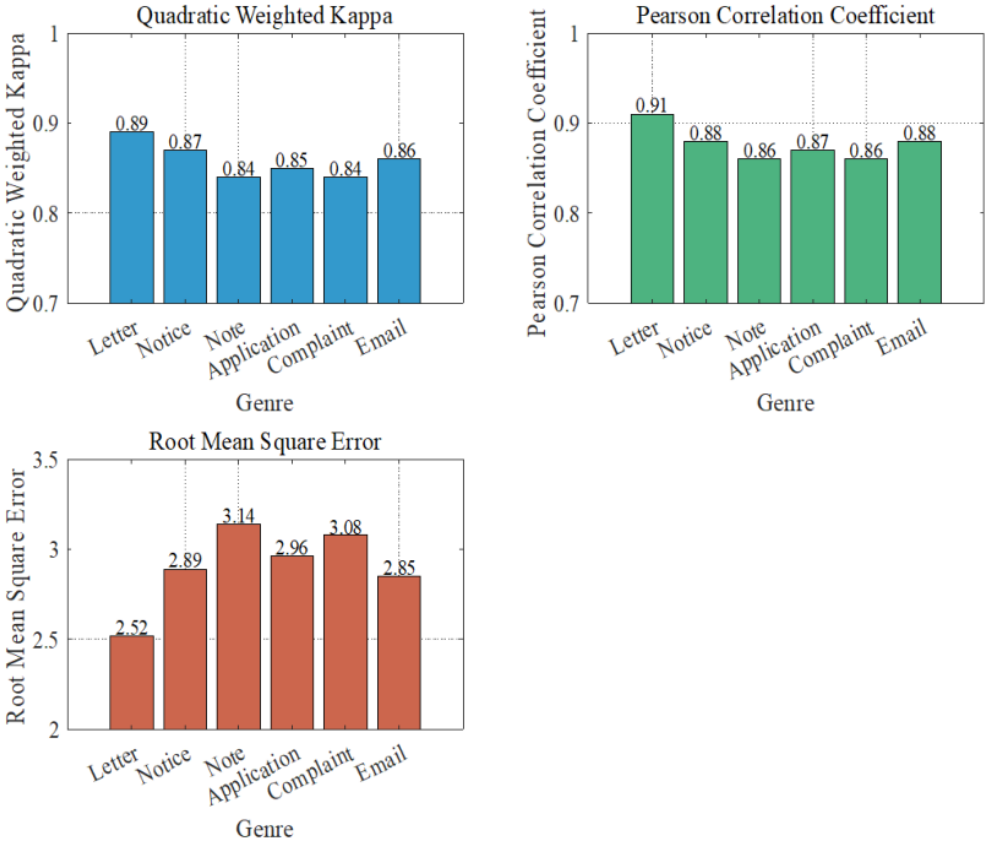

FIGURE 1. Performance evaluation results of the automatic scoring model

TABLE 1. Results of manual evaluation of diagnostic feedback quality

Feedback source	Content accuracy	Personalization	Practicality of teaching
Model generates feedback	4.32	4.18	4.25
Teacher feedback	4.58	4.35	4.62

feedback created by the model and control group students were not given feedback, just the scoring results.

The two groups were obliged to review their initial essays according to the information they have gotten when they received the feedback or the score. Two teachers of the texts were used to rate the revised texts after being revised, and the improvement score of each text (revised score-original score) was done and the improvement score of each text compared in the two groups. Table 2 presents the results of the experiment.

The findings indicate that average improvement score of the experimental group was 6.25 points and the average improvement score of control group was 2.18 points which means there was a big difference between the experimental group and the control group. This finding implies that diagnostic feedback proves to be more effective when it comes to the improvement of the writing of lower and middle score students, which in other words is able to assist them in discovering issues in writing and undertaking specific revisions.

The students who were in the control group were simply given the scoring result and were not given a specific advice on how to revise. The effect of their revisions was in fact limited to improve and some students even registered a decline in their scores. The reason behind this was that, with no definite guide, the revisions by students could end up creating new mistakes.

The synopsis of the re-writes on the experimental group students demonstrated that the higher enhancement was reflected in the content completeness with an average of 2.15 points; language accuracy with an average of 1.85 points; stylistic conformity with an average of 1.25 points, and structural clarity with an average of 1.00 points. This distribution has to deal with the specificity of the diagnostic feedback along various dimensions.

Evaluation of content completeness is usually very much

TABLE 2. Results of the experiment verifying the feedback improvement effect

Group	Sample size	Average score of original manuscript	Average score of revised manuscript	Average score increase
experimental group	60	70.32	76.57	6.25
control group	60	70.18	72.36	2.18

precise, indicating the content clearly where there are gaps in terms of elements or points of information that are not present in the content and as such these gaps can be very easily understood and corrected by students.

Advancements in the precision of language, however, entail several stages, such as grammar and vocabulary. Certain types of errors like the wrong use of tense and wrong collocation can only be corrected in case the student has a good understanding of the language.

When it comes to the particular contents of the feedback, the students tended to consider the suggestions about the improvement, made by the model, to be most useful, including the offered revision examples, which demonstrated graphically the right formulations. Other students have also reported that the model gave them specific locations and type of errors when they identified errors thus they were able to get the issues on the right track.

V. CONCLUSION

This study focuses on the scoring and feedback issues in vocational college English applied writing instruction. Based on a large language model, an automatic scoring and diagnostic feedback generation system was constructed. A corpus of vocational college English writing texts containing six applied writing genres was built, and a four-dimensional scoring standard consistent with the actual teaching situation in vocational colleges was established. Based on this, automatic scoring and diagnostic feedback generation modules were implemented.

Experimental results show that the average weighted consistency coefficient between the multi-dimensional automatic scoring model and human scoring reached 0.87. Diagnostic feedback received high evaluations in three dimensions: content accuracy, personalization, and teaching practicality. The feedback improvement effect verification experiment showed that students who received diagnostic feedback improved their scores by an average of 6.25 points, significantly higher than the control group who only received scoring results.

This study verifies the technical feasibility and teaching effectiveness of applying a large language model to the automatic scoring and diagnostic feedback generation of vocational college English applied writing. The study still has certain limitations, such as the accuracy of diagnostic feedback in generating improvement suggestions still falling short of the teacher's level, and the consistency of the model's scoring for some text genres needs further improvement.

Future research could focus on optimizing the model architecture, improving feedback generation strategies, and fusing multimodal information to further enhance the system's performance and its applicability for teaching, while also exploring ways to promote and apply the system in real-world teaching scenarios.

AUTHOR CONTRIBUTIONS

Lanna He designed, collected and analyzed the data, and drafted the manuscript. Lanna He conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Lanna He participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] T. Ismailia, "Teaching English for vocational college students by using a Project-based learning approach," Journal of English in Academic and Professional Communication, vol. 10, no. 1, pp. 19-35, 2024, doi: 10.25047/jeapco.v10i1.4546.

[2] A. Z. Idham, W. Rauf, and A. Rajab, "Navigating the transformative impact of artificial intelligence on English language teaching: Exploring challenges and opportunities," Jurnal Edukasi Saintifik, vol. 4, no. 1, pp. 8-14, 2024, doi: 10.56185/jes.v4i1.620.

[3] H. Husnaini, "Development of Self Esteem-Oriented Micro Teaching Materials for IAIN Palopo English Education Students," IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, vol. 10, no. 1, pp. 538-560, 2022, doi: 10.24256/ideas.v10i1.2408.

[4] M. H. Sawalmeh and M. Dey, "Globalization and the increasing demand for spoken English teachers," Research Journal in Advanced Humanities, vol. 4, no. 2, pp. 47-60, 2023, doi: 10.58256/rjah.v4i2.1097.

[5] Z. N. Ghafar, "English for specific purposes in English language teaching: design, development, and environment-related challenges: an overview," Canadian Journal of Language and Literature Studies, vol. 2, no. 6, pp. 32-42, 2022, doi: 10.53103/cjlls.v2i6.72.

[6] H. Z. Alghamdy, "English teachers' practice of classroom discourse in light of zone of proximal development theory and scaffolding techniques," Journal of Language Teaching and Research, vol. 15, no. 1, pp. 46-54, 2024, doi: 10.17507/jltr.1501.06.

[7] U. Lee, H. Jung, Y. Jeon, et al., "Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education," Education and Information Technologies, vol. 29, no. 9, pp. 11483-11515, 2024, doi: 10.1007/s10639-023-12249-8.

[8] A. S. Uyun, "Teaching English speaking strategies," Journal of English Language Learning, vol. 6, no. 1, pp. 14-23, 2022, doi: 10.31949/jell.v6i1.2475.

[9] S. N. Çelik and H. B. Memduhoglu, "The Evaluation of the English Language Teacher Education Program in Turkey," International Journal of Educational Methodology, vol. 8, no. 4, pp. 833-851, 2022, doi: 10.12973/ijem.8.4.833.

[10] M. Muhajir and A. S. A. Syamsu, "The Creative Exploitation of Pecha Kucha's Presentation Technique in English Teaching Classes," Qalam: Jurnal Ilmu Kependidikan, vol. 11, no. 2, pp. 67-71, 2022, doi: 10.33506/jq.v11i2.2010.

[11] M. J. Tividad, "The English language needs of a technical vocational institution," Psychology and Education: A Multidisciplinary Journal, vol. 16, no. 3, pp. 256-275, 2024, doi: 10.2139/ssrn.4695007.

[12] F. X. R. Baskara, "Integrating ChatGPT into EFL writing instruction: Benefits and challenges," International Journal of Education and Learning, vol. 5, no. 1, pp. 44-55, 2023, doi: 10.31763/ijele.v5i1.858.

[13] P. Bannister, A. S. Urbieta, and E. A. Peñalver, "A systematic review of generative AI and (English medium instruction) higher education," Aula Abierta, vol. 52, no. 4, pp. 401-409, 2023, doi: 10.17811/rifie.52.4.2023.401-409.

[14] Y. C. Tseng and Y. H. Lin, "Enhancing English as a Foreign Language (EFL) Learners' Writing with ChatGPT: A University-Level Course Design," Electronic Journal of e-Learning, vol. 22, no. 2, pp. 78-97, 2024, doi: 10.34190/ejel.21.5.3329.