AUTHOR-ACCEPTED MANUSCRIPT

Automated Writing Evaluation with Corrective Feedback

Mengyang Qiu, Izia Xiaoxiao Wang, Tran Minh Nguyen,
Zihao Huang, Yang Gu, Qingyu Gao, Zelong Li,
Siliang Liu, and Jungyeul Park

IEEE copyright notice

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Published version

Qiu, M., Wang, I. X., Nguyen, T. M., Huang, Z., Gu, Y., Gao, Q., Li, Z., Liu, S., & Park, J. (2026). Automated writing evaluation with corrective feedback. *2026 International Conference on Artificial Intelligence and Education (ICAIE)*, 142-146. https://doi.org/10.1109/ICAIE70080.2026.11751221

IEEE Xplore publication record

https://ieeexplore.ieee.org/document/11751221

Automated Writing Evaluation with Corrective Feedback

Mengyang Qiu*
Department of Speech-Language Pathology
Saint Elizabeth University
Morristown, NJ, USA
mqiu@steu.edu

Izia Xiaoxiao Wang
Zurich Center for Linguistics
University of Zurich
Zurich, Switzerland
ziya.x.wang@gmail.com

Tran Minh Nguyen
Open Writing Evaluation
Côtes d'Amor
France
nguyentranminh@gmail.com

Zihao Huang
Open Writing Evaluation
Côtes d'Amor
France
hzhhzh1999@gmail.com

Yang Gu
Open Writing Evaluation
Côtes d'Amor
France
yangguamy@gmail.com

Qingyu Gao
Open Writing Evaluation
Côtes d'Amor
France
chrisgaoqy@gmail.com

Zelong Li
Department of Computer Science
University College London
London, UK
zelong.li.xdd@gmail.com

Siliang Liu
Open Writing Evaluation
Côtes d'Amor
France
natashanatasha625@gmail.com

Jungyeul Park*
Culture Technology Research Institute
KAIST
Daejeon, South Korea
jungyeul@kaist.ac.kr

***Abstract*—The use of technology in second language writing assessment has become ubiquitous. In particular, automated writing evaluation (AWE) and grammatical error correction (GEC) are popular tools for providing immediate, individualized feedback to learners. Traditionally, AWE (automated scoring) and GEC (automated correction) have been developed separately. In this paper, we present an integrated system that combines neural AWE and GEC to deliver both an objective essay score and corrective feedback to a learner in one go. The system allows learners to simulate an essay-writing test by submitting an essay and receiving an automated score along with suggested grammatical corrections. This integration is intended to bridge the gap between scoring and feedback, alleviating the burden on instructors by handling large volumes of essays more efficiently. A demonstration of the system is available at https://youtu.be/CoqmBqqcXio.**

***Index Terms*—automated writing evaluation, grammatical error correction, corrective feedback, essay scoring, L2 writing**

I. Introduction

Second language (L2) writing instructors are often expected to provide extensive feedback on student essays, but doing so for every student draft is time-consuming and labor-intensive [1]. Automated writing evaluation (AWE)[1] and grammatical error correction (GEC) tools offer potential solutions by reducing instructor workload and giving learners instant feedback. These systems align with the need for more efficient writing practice in the digital age, providing immediate scores and corrections that help students improve iteratively.

AWE is defined as "the process of evaluating and scoring written prose via computer programs" [2]. In essence, an AWE system predicts a holistic score or a set of trait scores for an essay, approximating human grading. Research has shown that scoring essays across prompts (i.e. training a model on multiple writing prompts) and across multiple writing traits can provide more robust and comprehensive assessments than prompt-specific scoring [3]. We follow this approach, arguing that an effective AWE component should produce analytic scores along several rubrics (e.g. content, organization, etc.) rather than only a single holistic score.

GEC is the task of automatically detecting and correcting grammatical and writing errors in text. As the number of L2 English learners grows and the demand for timely feedback increases, GEC has become highly popular in both NLP research and educational applications. State-of-the-art GEC systems are often based on sequence-to-sequence neural models that transform a learner's original text into a corrected text [4]. These systems can provide explicit corrective feedback, highlighting errors and suggesting corrections, which is invaluable for language learners.

While AWE and GEC have typically been developed as separate tools, we propose a unified system that integrates both scoring and correction into a single platform. This integration allows learners to receive holistic and analytic scores alongside explicit grammatical corrections, offering a more comprehensive form of feedback. By simulating real testing

*Corresponding authors.

[1]We use AWE to refer broadly to systems that automatically score and/or provide feedback on writing. This encompasses what is also called automated essay scoring (AES) when referring specifically to score prediction. In this paper, the terms are used interchangeably.

conditions, where a student writes an essay and receives both an evaluation and suggested revisions, our system supports both self-assessment and iterative improvement.

What sets our approach apart is its dual focus: 1) providing interpretable, rubric-based scores across multiple writing traits, and 2) delivering in-line corrective feedback that highlights and explains grammatical errors. The two components work together to offer a richer learning experience than either system alone. In this paper, we describe the system architecture, underlying models, and interface (Sections II and III), and discuss future directions (Section IV). Although our current implementation focuses on English L2 writing, the design is modular and can be extended to other languages and educational contexts.

II. System Architecture and Methodology

Our system consists of two core NLP components, an AWE model and a GEC model, integrated into a web-based platform that learners can interact with. Figure 1 illustrates the overall workflow. In implementation, the two components operate in tandem: when a user submits an essay, the text is sent to both the scoring module and the correction module. The outputs are then combined to generate a detailed feedback report for the user. Below, we describe each component and the techniques we employed to improve their performance.

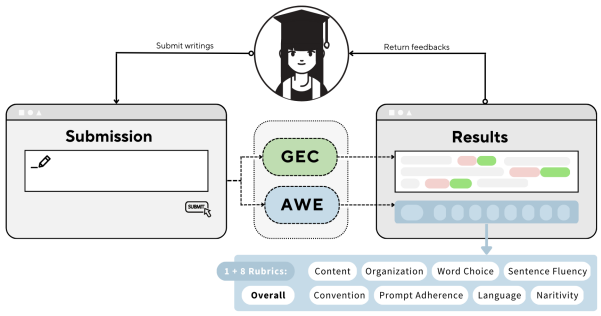


Fig. 1. System workflow of the integrated AWE+GEC system. A student writes an essay in the submission interface and clicks "Submit". The system processes the text through AWE and GEC models. The GEC module produces a corrected version of the essay (highlighting edits), and the AWE module produces a set of scores. This design allows learners to practice writing and immediately see both what errors to "correct" and how they might be "scored" under exam criteria, effectively simulating an exam scenario.

a) AWE: For automated scoring, we build on prior work in neural essay scoring using the ASAP dataset and the enhanced ASAP++ extension [5] for training our scoring models. The ASAP dataset consists of eight sets of essays written to eight different prompts (source topics), each essay with one or more human scores (usually a holistic score, and for some prompts additional trait scores such as content, organization, etc.). ASAP++ provides additional rubric-based scores for the prompts that originally had only holistic scores, yielding a richer set of multi-trait annotations. In total, we leverage eight writing prompts worth of data, each essay annotated with three to five trait scores (e.g., content, organization, word choice) and an overall score.

For the current system, we use a denoised version of the ASAP dataset, cleaned using simple prompting techniques with `gpt-3.5` to correct encoding artifacts and anonymized tokens[2]. The original dataset applied named entity recognition via Stanford CoreNLP [7], replacing entities with placeholders like `@PERSON1`. Additionally, over 5% of sentences contain encoding issues or non-word entities. We classify these as noise and apply a two-step denoising process using `gpt-3.5-turbo-instruct`: one prompt fixes encoding errors, and the other replaces anonymized tokens with generic names. Because `gpt-3.5` may overcorrect grammar, we restore the original structure using `.m2` annotations generated by `ERRANT` [8], preserving only the intended encoding and token substitutions.

After denoising, we train a `roberta-base` model [9] to predict trait scores using linear regression. Training uses the default settings from the Hugging Face `Trainer` class,[3] with scores normalized to [0, 1] and scaled by 100 for Quadratic Weighted Kappa (QWK) evaluation using the standard script from `ets.org`. The average QWK across prompts increased from 0.6047 to 0.6143 after denoising—an improvement of about 1 percentage point. While this gain may appear modest in aggregate, it is consistent: six of the eight prompts show improvement, with Prompt 1 exhibiting a substantial gain of over 11 points (0.6187 to 0.7344). The improvement is also reflected in reduced prediction error, with both Mean Squared Error and Root Mean Square Deviation decreasing after denoising [6], suggesting a genuine, if incremental, effect of denoising on scoring quality. Table I shows a prompt-by-prompt breakdown.

TABLE I
PROMPT-BY-PROMPT QWK RESULTS: E.G. `PROMPT 1` REPRESENTS THAT PROMPT 1 IS USED AS A TEST SET, AND PROMPTS 2-7 AS A TRAINING SET, AND PROMPT 8 AS A DEV SET.

	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7	Prompt 8	Average
Original	0.6187	0.6308	0.6962	0.6626	0.7158	0.5924	0.4622	0.4586	0.6047
Cleaned	0.7344	0.5485	0.6978	0.6540	0.7416	0.6156	0.4637	0.4589	0.6143

b) GEC: Our GEC module is based on the BERT-fused encoder-decoder model proposed by [10], implemented using the `fairseq` toolkit [11]. The model combines a Transformer encoder-decoder architecture with contextual embeddings from a fine-tuned BERT model on standard English learner corpora from the *BEA-2019 Shared Task* [12], including FCE, NUCLE, Lang-8, and W&I+LOCNESS. All the incorrect sentences are spell-checked and added back in order to augment the size of the original training set ($1,157,324$ sentence pairs), resulting in $1,718,693$ sentence pairs in total. $6,575$ sentence pairs from the FCE and W&I+LOCNESS development sets are used as validation. The model achieves an $F_{0.5}$ score of 65.29 on the *BEA-2019* test set.

In the current system, we use only the BERT-based model. However, our framework is modular and supports other off-

[2]This denoising procedure was previously presented in [6].

[3]https://huggingface.co/docs/transformers/main_classes/trainer

the-shelf implementations, such as those based on BART [13] and T5 [14], allowing for easy switching or ensembling in future versions. Another design decision is that we do not rely on large language models (LLMs) like GPT-4 for GEC in our system. While these models are capable of fluent rewriting, recent work has shown that they tend to over-correct, often violating the "minimal edits" principle expected in learner-facing feedback [15], [16].

In particular, our proficiency-based evaluation [16] found that for prompting-based LLMs, overcorrection is especially pronounced at the advanced (CEFR C) proficiency level. Fine-tuning reduces overcorrection but at the cost of degraded precision for advanced-level writing. In both cases, dedicated GEC models continue to outperform prompting-based and fine-tuned LLM approaches in $F_{0.5}$ across all proficiency levels. In addition, relying on API-based LLMs raises concerns about cost, latency, and data privacy. For these reasons, our current implementation uses a fine-tuned, open-source model that enables more controllable correction while remaining efficient and locally deployable. We are actively exploring controlled and hybrid integration strategies that leverage LLM capabilities while preserving the minimal-edit behavior essential for corrective feedback.

III. Demo Interface and Features

The system provides users with a unified interface to submit essays, view automated corrections, and receive trait-based scores. Once a student submits an essay, the platform displays both inline grammatical corrections and a scoring breakdown.

To present feedback clearly, we align edits using `ERRANT` [8], and highlight them directly in the learner's original text, as shown in Figure 2. Red highlights indicate deletions (e.g., incorrect words or phrases), while green highlights mark insertions or corrections. By preserving the original text with markup, we ensure the feedback is contextual–the student sees where and what the mistake was, rather than just seeing a corrected version in isolation. The goal is to provide intuitive feedback that makes error detection and revision easier. In future versions, we plan to supplement these highlights with brief explanations for common error types (e.g., subject–verb agreement or verb tense), but the current system focuses on clean, direct visual feedback.

Below the annotated essay, the interface lists the automated scores. In the example screenshot, the system gave an Overall Score of 12 (out of 100) for the essay, indicating a very low writing proficiency. In this case, the essay was very short and had many grammatical errors, so the scoring reflects that. These scores simulate how a human grader might rate the essay using a rubric. It is important to note that the score ranges are adjustable. Our demonstration uses the percentage style for simplicity. If an instructor prefers the traditional 1–5 scale for each trait, the system could be configured to show that instead.

We also allow the possibility of normalizing scores to a user's proficiency level. For instance, if a beginner and an advanced learner both use the system, an instructor might want to normalize or interpret the scores relative to typical performance at those levels (so a score of 50 for a beginner might be a good score, whereas for an advanced learner it would be poor). Currently, we leave this interpretation to the users/instructors, but the data could be used to create percentile norms in the future.

From a user experience perspective, the system provides immediate results (processing time is usually only a few seconds for a short essay). This immediacy means a student can write a practice essay under test-like conditions and get feedback right away, enabling a cycle of practice and improvement. Beyond the student-facing features, the system also has an instructor mode (planned for future development) where a teacher can log in and see the results from their class.

In the current prototype, we have a basic admin interface that allows an instructor to view a submitted essay and the system's feedback. We are exploring an instructor intervention feature (see Figure 3) where the instructor can optionally review and adjust the automated feedback before it is returned to the student. In that scenario, the system would generate a default set of scores and corrections, then the instructor could modify any part (for example, override a score if it seems inaccurate, or add an additional comment for the student). The adjusted feedback would then be sent to the student, bridging automated and human evaluation. While this is not fully implemented in the current demo, it represents a potential future extension to increase teacher control and trust in the system's outputs.



Fig. 2. Screenshot of the user interface. Red highlighted segments (prefixed "–") indicate deletions; green segments (prefixed "+") indicate corrections. Below the essay text, the interface displays both overall and trait-specific AWE scores.

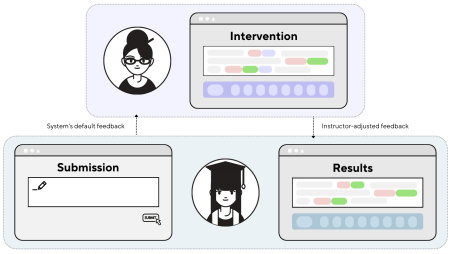


Fig. 3. Future instructor intervention workflow. After a student submits an essay, the system generates default automated feedback. A teacher can review, adjust, or override the system's feedback, including both GEC and AWE outputs, before it is delivered to the student.

IV. Discussion and Future Perspectives

Our integrated AWE+GEC system demonstrates the feasibility and benefits of combining automated scoring with automated corrective feedback for English writing. By evaluating content and form together, the system provides a more holistic feedback loop to learners. There are, however, several avenues for future improvement.

Currently, the system's feedback is limited to grammatical corrections and numeric scores. It does not generate explanatory comments or suggestions beyond pointing out errors. In contrast, human teachers often give more detailed comments (e.g. "Your introduction is clear, but you need to develop your argument more in the body paragraphs."). In the future, we plan to extend the system to provide textual feedback in addition to scores–for example, using an LLM to generate a brief summary of the essay's strengths and weaknesses [17]. This could make the feedback even more instructive.

Of course, the ultimate success of an educational tool depends on user trust. Some prior studies have examined the effectiveness of AWE on improving learners' writing performances and students' perceptions of AWE systems, for example, whether they find the automated feedback fair, accurate, and helpful [18]. We recognize that not all users will immediately trust a computer-generated score or correction.

Preliminary informal feedback from a small group of beta testers has been encouraging, particularly regarding the immediacy of the feedback and the clarity of the error highlights. However, these observations are anecdotal. We therefore plan to conduct a user study to evaluate instructor and student perceptions of the system's feedback. Specifically, we will survey participants on factors such as perceived fairness of the grading, clarity of the feedback, and willingness to use such a tool regularly (see Table II for sample survey questions). By collecting these data, we hope to identify aspects of the system that may require further refinement. For example, if users consistently distrust the scoring, we may need to improve its explanations or calibration. In addition to perception data, we also plan to evaluate the system's real-world effectiveness by comparing its scores and corrections with human ratings and feedback. We intend to report these findings in future work and use them to further refine the user experience.

Another important limitation is that the current version of the system focuses exclusively on English learner writing. While this allows for targeted modeling using established resources such as ASAP++ and the *BEA-2019* datasets, it limits the system's applicability to learners of other languages. Fortunately, many of the underlying neural techniques are language-agnostic and can be adapted given appropriate training data. In recent years, there has been growing availability of GEC datasets for European languages [19] and non-European languages, such as Chinese [20] and Korean [21]. In parallel, the `ERRANT` annotation framework has been extended to multiple languages [22]–[24]. For the AWE component, a key challenge in multilingual extension lies in the availability of scored essay corpora. Although there are ongoing efforts to develop annotated writing corpora in various languages, few are as richly labeled with multi-trait rubrics as ASAP++. One promising direction is the use of universal scoring rubrics that evaluate traits like content, organization, and coherence in a language-neutral way, which would allow cross-lingual transfer learning or fine-tuning while preserving trait consistency. We aim to explore these multilingual extensions in future work and evaluate the feasibility of adapting our platform for additional languages, making the system a more inclusive educational tool for global learning contexts.

TABLE II
Questionnaire for the perception experiment of the functionality (GEC+AWE) and effectiveness of the current system

Perceived fairness and accuracy of GEC and AWE:
- On a scale of 1–5, how fair do you think the grammar feedback on this writing to be? (1 = not fair at all, 5 = very fair)
- On a scale of 1–5, how accurate and trustworthy do you think the grammar correction on this writing to be? (1 = not accurate/trustworthy at all, 5 = very accurate/trustworthy)
- On a scale of 1–5, how fair do you think the grading and evaluation of this writing to be? (1 = not fair at all, 5 = very fair)
- On a scale of 1–5, how accurate and trustworthy do you think the grading and evaluation of this writing to be? (1 = not accurate/trustworthy at all, 5 = very accurate/trustworthy)

Perceived clarity and helpfulness:
- On a scale of 1–5, how clear do you think the grading and feedback on this writing to be? (1 = not clear at all, 5 = very clear)
- On a scale of 1–5, how helpful do you perceive the feedback on this writing to be in improving this student's writing skills? (1 = not helpful at all, 5 = very helpful)
- On a scale of 1–5, how helpful do you think rubrics with criteria and scores to be in understanding the grading and feedback on this writing assignment? (1 = not helpful at all, 5 = very helpful)

Mental effort:
- On a scale of 1–5, how difficult do you find it to perceive the grading and feedback on this writing to be?
- On a scale of 1–5, how much mental effort does this task (differ for stages 1 and 2) require from you to perceive the feedback?

Attitudinal responses:
- On a scale of 1–5, how comfortable do you feel about perceiving/using this system to receive the grading and feedback on this writing?
- On a scale of 1–5, how effective do you feel about perceiving/using this system to receive the grading and feedback on this writing?

V. Conclusion

We have presented an integrated system that brings together automated essay scoring and grammatical error correction in a single platform aimed at second language writing practice. The system, currently implemented for English, offers instantaneous scoring along multiple dimensions and explicit corrective feedback that traditionally required significant teacher effort. By providing a way for learners to practice writing with the safety net of immediate feedback, we hope to enhance the learning experience and outcomes in L2 writing instruction. Our approach balances the strengths of neural models with practical considerations for educational deployment, and we believe it can serve as a foundation for richer AI-driven writing tutors.

REFERENCES

[1] M. Cho, "The Use and Efficacy of Correction Symbols in KFL Writing," *The Korean Language in America*, vol. 22, no. 1, pp. 83–92, 2018. [Online]. Available: https://www.jstor.org/stable/10.5325/korelangamer.22.1.0083

[2] M. D. Shermis and J. C. Burstein, "Introduction," in *Automated Essay Scoring: A Cross-disciplinary Perspective*, M. D. Shermis and J. C. Burstein, Eds. New York: Routledge, 2003, pp. xiii–xvii.

[3] R. Ridley, L. He, X.-y. Dai, S. Huang, and J. Chen, "Automated Cross-prompt Scoring of Essay Traits," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, pp. 13 745–13 753, 5 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17620

[4] C. Bryant, Z. Yuan, M. R. Qorib, H. Cao, H. T. Ng, and T. Briscoe, "Grammatical Error Correction: A Survey of the State of the Art," *Computational Linguistics*, vol. 49, no. 3, pp. 643–701, 2023. [Online]. Available: https://doi.org/10.1162/coli_a_00478

[5] S. Mathias and P. Bhattacharyya, "ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 5 2018, pp. 1169–1173. [Online]. Available: https://aclanthology.org/L18-1187

[6] J. Park and M. Qiu, "Frustratingly Simple Prompting-based Text Denoising," in *The Second Tiny Papers Track at ICLR 2024*, 2024, pp. 1–5. [Online]. Available: https://openreview.net/forum?id=XlJRjhIkNi

[7] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, K. Bontcheva and J. Zhu, Eds. Baltimore, Maryland: Association for Computational Linguistics, 6 2014, pp. 55–60. [Online]. Available: https://aclanthology.org/P14-5010

[8] C. Bryant, M. Felice, and T. Briscoe, "Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, 7 2017, pp. 793–805. [Online]. Available: https://aclanthology.org/P17-1074

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *CoRR*, vol. abs/1907.1, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[10] M. Kaneko, M. Mita, S. Kiyono, J. Suzuki, and K. Inui, "Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 7 2020, pp. 4248–4254. [Online]. Available: https://aclanthology.org/2020.acl-main.391

[11] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A Fast, Extensible Toolkit for Sequence Modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, 6 2019, pp. 48–53. [Online]. Available: https://aclanthology.org/N19-4009

[12] C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe, "The BEA-2019 Shared Task on Grammatical Error Correction," in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florence, Italy: Association for Computational Linguistics, 8 2019, pp. 52–75. [Online]. Available: https://aclanthology.org/W19-4406

[13] S. Katsumata and M. Komachi, "Stronger Baselines for Grammatical Error Correction Using a Pretrained Encoder-Decoder Model," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 12 2020, pp. 827–832. [Online]. Available: https://aclanthology.org/2020.aacl-main.83

[14] S. Rothe, J. Mallinson, E. Malmi, S. Krause, and A. Severyn, "A Simple Recipe for Multilingual Grammatical Error Correction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, 8 2021, pp. 702–707. [Online]. Available: https://aclanthology.org/2021.acl-short.89

[15] T. Fang, S. Yang, K. Lan, D. F. Wong, J. Hu, L. S. Chao, and Y. Zhang, "Is ChatGPT a Highly Fluent Grammatical Error Correction System? A Comprehensive Evaluation," pp. 1–16, 2023. [Online]. Available: https://arxiv.org/abs/2304.01746

[16] M. Zeng, J. Kuang, M. Qiu, J. Song, and J. Park, "Evaluating Prompting Strategies for Grammatical Error Correction Based on Language Proficiency," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italy: ELRA and ICCL, 5 2024, pp. 6426–6430. [Online]. Available: https://aclanthology.org/2024.lrec-main.569

[17] M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth, "Exploring LLM prompting strategies for joint essay scoring and feedback generation," in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 283–298. [Online]. Available: https://aclanthology.org/2024.bea-1.23/

[18] M. S. A. Ariyanto, N. Mukminatien, and S. Tresnadewi, "College Students' Perceptions of an Automated Writing Evaluation as a Supplementary Feedback Tool in a Writing Class," *Jurnal Ilmu Pendidikan*, vol. 27, no. 1, pp. 41–51, 2021.

[19] A. Masciolini, A. Caines, O. D. Clercq, J. Kruijsbergen, M. Kurfalı, R. M. Sánchez, E. Volodina, and R. Östling, "The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL," in *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, R. M. Sánchez, D. Alfter, E. Volodina, and J. Kallas, Eds. Tallinn, Estonia: University of Tartu Library, Mar. 2025, pp. 1–33. [Online]. Available: https://aclanthology.org/2025.nlp4call-1.1/

[20] Y. Zhang, Z. Li, Z. Bao, J. Li, B. Zhang, C. Li, F. Huang, and M. Zhang, "MuCGEC: a Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 7 2022, pp. 3118–3130. [Online]. Available: https://aclanthology.org/2022.naacl-main.227

[21] S. Yoon, S. Park, G. Kim, J. Cho, K. Park, G. T. Kim, M. Seo, and A. Oh, "Towards standardizing Korean Grammatical Error Correction: Datasets and Annotation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, 7 2023, pp. 6713–6742. [Online]. Available: https://aclanthology.org/2023.acl-long.371

[22] R. Belkebir and N. Habash, "Automatic Error Type Annotation for Arabic," in *Proceedings of the 25th Conference on Computational Natural Language Learning*, A. Bisazza and O. Abend, Eds. Online: Association for Computational Linguistics, 11 2021, pp. 596–606. [Online]. Available: https://aclanthology.org/2021.conll-1.47

[23] Y. Gu, Z. Huang, M. Zeng, M. Qiu, and J. Park, "Improving Automatic Grammatical Error Annotation for Chinese Through Linguistically-Informed Error Typology," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, 1 2025, pp. 2781–2798. [Online]. Available: https://aclanthology.org/2025.coling-main.189/

[24] M. Qiu, T. M. Nguyen, Z. Huang, Z. Li, Y. Gu, Q. Gao, S. Liu, and J. Park, "Multilingual grammatical error annotation: Combining language-agnostic framework with language-specific flexibility," in *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*. Vienna, Austria: Association for Computational Linguistics, 2025, pp. 202–212. [Online]. Available: https://aclanthology.org/2025.bea-1.15/