References

1. Binhammad, M.H.Y., Othman, A., Abuljadayel, L., Al Mheiri, H., Alkaabi, M., Almarri, M.: Investigating how generative ai can create personalized learning materials tailored to individual student needs. Creative Education **15**(7), 1499–1523 (2024)
2. Burstein, J., Chodorow, M., Leacock, C.: Automated essay evaluation: The criterion online writing service. Ai magazine **25**(3), 27–27 (2004)
3. Chen, B., Zhang, Z., Langrené, N., Zhu, S.: Unleashing the potential of prompt engineering in large language models: a comprehensive review. arXiv preprint arXiv:2310.14735 (2023)
4. Chen, E., Wang, D., Xu, L., Cao, C., Fang, X., Lin, J.: A systematic review on prompt engineering in large language models for k-12 stem education. arXiv preprint arXiv:2410.11123 (2024)
5. Chen, H., He, B.: Automated essay scoring by maximizing human-machine agreement. In: Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 1741–1752 (2013)
6. Crossley, S.A., Tian, Y., Baffour, P., Franklin, A., Benner, M., Boser, U.: A large-scale corpus for assessing written argumentation: Persuade 2.0. Assessing Writing **61**, 100865 (2024)
7. Deane, P.: The importance of assessing student writing and improving writing instruction. research notes. Educational Testing Service (2022)
8. Doewes, A., Kurdhi, N., Saxena, A.: Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In: 16th International Conference on Educational Data Mining, EDM 2023. pp. 103–113. International Educational Data Mining Society (IEDMS) (2023)
9. Dong, F., Zhang, Y., Yang, J.: Attention-based recurrent convolutional neural network for automatic essay scoring. In: Proceedings of the 21st conference on computational natural language learning (CoNLL 2017). pp. 153–162 (2017)
10. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al.: Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858 (2022)
11. García-Méndez, S., de Arriba-Pérez, F., Somoza-López, M.d.C.: A review on the use of large language models as virtual tutors. Science & Education pp. 1–16 (2024)
12. Jones, S., Myhill, D.: Discourses of difference? examining gender differences in linguistic characteristics of writing. Canadian Journal of Education/Revue canadienne de l'éducation pp. 456–482 (2007)
13. Kwako, A., Ormerod, C.: Can language models guess your identity? analyzing demographic biases in ai essay scoring. In: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024). pp. 78–86 (2024)
14. Lagakis, P., Demetriadis, S.: Automated essay scoring: A review of the field. In: 2021 International Conference on Computer, Information and Telecommunication Systems (CITS). pp. 1–6. IEEE (2021)
15. Lee, M., Liang, P., Yang, Q.: Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In: Proceedings of the 2022 CHI conference on human factors in computing systems. pp. 1–19 (2022)
16. Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., Chen, G.: Can large language models write reflectively. Computers and Education: Artificial Intelligence **4**, 100140 (2023)

17. Litman, D., Zhang, H., Correnti, R., Matsumura, L.C., Wang, E.: A fairness evaluation of automated methods for scoring text evidence usage in writing. In: International Conference on Artificial Intelligence in Education. pp. 255–267. Springer (2021)
18. Loukina, A., Madnani, N., Zechner, K.: The many dimensions of algorithmic fairness in educational applications. In: Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications. pp. 1–10 (2019)
19. Mayfield, E., Black, A.W.: Should you fine-tune bert for automated essay scoring? In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 151–162 (2020)
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
21. Olea, C., Tucker, H., Phelan, J., Pattison, C., Zhang, S., Lieb, M., White, J.: Evaluating persona prompting for question answering tasks. In: Proceedings of th e 10th international conference on artificial intelligence and soft computing, Sydney, Australia (2024)
22. Rodriguez, P.U., Jafari, A., Ormerod, C.M.: Language models and automated essay scoring. arXiv preprint arXiv:1909.09482 (2019)
23. Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A.: A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927 (2024)
24. Schaller, N.J., Ding, Y., Horbach, A., Meyer, J., Jansen, T.: Fairness in automated essay scoring: A comparative analysis of algorithms on german learner essays from secondary education. In: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024). pp. 210–221 (2024)
25. Sha, L., Li, Y., Gasevic, D., Chen, G.: Bigger data or fairer data?: augmenting bert via active sampling for educational text classification. In: International Conference on Computational Linguistics 2022. pp. 1275–1285. Association for Computational Linguistics (ACL) (2022)
26. Stahl, M., Biermann, L., Nehring, A., Wachsmuth, H.: Exploring llm prompting strategies for joint essay scoring and feedback generation. arXiv preprint arXiv:2404.15845 (2024)
27. Taghipour, K., Ng, H.T.: A neural approach to automated essay scoring. In: Proceedings of the 2016 conference on empirical methods in natural language processing. pp. 1882–1891 (2016)
28. de Vassimon Manela, D., Errington, D., Fisher, T., van Breugel, B., Minervini, P.: Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 2232–2242 (2021)
29. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
30. Williamson, D.M., Xi, X., Breyer, F.J.: A framework for evaluation and use of automated scoring. Educational measurement: issues and practice **31**(1), 2–13 (2012)
31. Xiao, C., Ma, W., Song, Q., Xu, S.X., Zhang, K., Wang, Y., Fu, Q.: Human-ai collaborative essay scoring: A dual-process framework with llms. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference. pp. 293–305 (2025)

32. Yan, L., Greiff, S., Teuber, Z., Gašević, D.: Promises and challenges of generative artificial intelligence for human learning. Nature Human Behaviour **8**(10), 1839–1850 (2024)
33. Yancey, K.P., Laflair, G., Verardi, A., Burstein, J.: Rating short l2 essays on the cefr scale with gpt-4. In: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). pp. 576–584 (2023)
34. Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., Chen, G.: Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 22466–22474 (2024)
35. Yang, R., Cao, J., Wen, Z., Wu, Y., He, X.: 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. pp. 1560–1569 (2020)
36. Yoon, H.J.: Interactions in efl argumentative writing: Effects of topic, l1 background, and l2 proficiency on interactional metadiscourse. Reading and Writing **34**(3), 705–725 (2021)
37. Yoshida, L.: The impact of example selection in few-shot prompting on automated essay scoring using gpt models. In: International Conference on Artificial Intelligence in Education. pp. 61–73. Springer (2024)
38. Zesch, T., Wojatzki, M., Scholten-Akoun, D.: Task-independent features for automated essay grading. In: Proceedings of the tenth workshop on innovative use of NLP for building educational applications. pp. 224–232 (2015)